2013

SCIENTOMETRICS

14th International Society of Scientometrics and Informetrics Conference

15th – 19th July 2013

Vienna, Austria

PROCEEDINGS Volume II

PROCEEDINGS OF
ISSI 2013
Vienna

VOLUME 2

14th International Society of
Scientometrics and Informetrics Conference

Vienna, Austria
15th to 20th July 2013

Editors

Juan Gorraiz, Edgar Schiebel, Christian Gumpenberger, Marianne Hörlesberger, Henk Moed

Sponsors

ASIS&T, USA
Elsevier B.V.
EBSCO Information Services, USA
Federal Ministry for Science and Research, Austria
Federal Ministry for Transport, Innovation and Technology, Austria
Information Assistant, Verein für Informationsmanagement, Vienna
ORCID, Inc.
Science-Metrix/R&D Reports
Swets Information Services
Thomson Reuters
ZSI - Centre for Social Innovation, Vienna

Printed by Facultas Verlags- und Buchhandels AG,
Stolbergasse 26, A-1050 Wien

ISBN: 978-3-200-03135-7
ISSN: 2175-1935

INDEX

KEYNOTE 1

SOCIAL NETWORK ANALYSIS 3

ORAL PRESENTATIONS 5

ACADEMIC CAREER STRUCTURES – HISTORICAL OVERVIEW GERMANY 1850-2013 7

ACADEMIC RESEARCH PERFORMANCE EVALUATION IN BUSINESS AND MANAGEMENT USING JOURNAL QUALITY CITING METHODOLOGIES 22

ACCESS TO UNIVERSITIES' PUBLIC KNOWLEDGE: WHO'S MORE REGIONALIST? 36

ADVANTAGES OF EVALUATING JOURNALS THROUGH ACCA - LIS JOURNALS (RIP[111]) 58

ANALYSIS OF JOURNAL IMPACT FACTOR RESEARCH IN TIME: DEVELOPMENT OF A SPECIALTY? 66

THE ANALYSIS OF RESEARCH THEMES OF OPEN ACCESS IN CHINA: IN THE PERSPECTIVE OF STRATEGIC DIAGRAM (RIP) 77

ANALYSIS OF THE WEB OF SCIENCE FUNDING ACKNOWLEDGEMENT INFORMATION FOR THE DESIGN OF INDICATORS ON 'EXTERNAL FUNDING ATTRACTION' 84

ANALYZING THE CITATION CHARACTERISTICS OF BOOKS: EDITED BOOKS, BOOK SERIES AND TYPES OF PUBLISHERS IN THE BOOK CITATION INDEX 96

THE APPLICATION OF CITATION-BASED PERFORMANCE CLASSES TO THE DISCIPLINARY AND MULTIDISCIPLINARY ASSESSMENT IN NATIONAL COMPARISON 109

APPROACH TO IDENTIFY SCI COVERED PUBLICATIONS WITHIN NON-PATENT REFERENCES IN PATENTS 123

ARE CITATIONS A COMPLETE MEASURE FOR THE IMPACT OF E-RESEARCH INFRASTRUCTURES? 136

ARE LARGER EFFECT SIZES IN EXPERIMENTAL STUDIES GOOD PREDICTORS OF HIGHER CITATION RATES? A BAYESIAN EXAMINATION. 152

[111] Research in progress paper

ARE THERE INTER-GENDER DIFFERENCES IN THE PRESENCE OF AUTHORS, COLLABORATION PATTERNS AND IMPACT? (RIP) 167

ASSESSING INTERNATIONAL COOPERATION IN S&T THROUGH BIBLIOMETRIC METHODS (RIP) 175

ASSESSING OBLITERATION BY INCORPORATION IN A FULL-TEXT DATABASE: JSTOR AND THE CONCEPT OF “BOUNDED RATIONALITY.” 185

ASSESSING THE MENDELEY READERSHIP OF SOCIAL SCIENCES AND HUMANITIES RESEARCH 200

ASSOCIATION BETWEEN QUALITY OF CLINICAL PRACTICE GUIDELINES AND CITATIONS GIVEN TO THEIR REFERENCES 215

AUTHOR NAME CO-MENTION ANALYSIS: TESTING A POOR MAN'S AUTHOR CO-CITATION ANALYSIS METHOD (RIP) 229

BIBLIOGRAPHIC COUPLING AND HIERARCHICAL CLUSTERING FOR THE VALIDATION AND IMPROVEMENT OF SUBJECT-CLASSIFICATION SCHEMES 237

BUILDING A MULTI-PERSPECTIVE SCIENTOMETRIC APPROACH ON TENTATIVE GOVERNANCE OF EMERGING TECHNOLOGIES 251

CAREER AGING AND COHORT SUCCESSION IN THE SCHOLARLY ACTIVITIES OF SOCIOLOGISTS: A PRELIMINARY ANALYSIS (RIP) 264

CITATION IMPACT PREDICTION OF SCIENTIFIC PAPERS BASED ON FEATURES 272

CITATION IMPACTS REVISITED: HOW NOVEL IMPACT MEASURES REFLECT INTERDISCIPLINARITY AND STRUCTURAL CHANGE AT THE LOCAL AND GLOBAL LEVEL 285

THE *CITER-SUCCESS-INDEX*: AN INDICATOR TO SELECT A SUBSET OF ELITE PAPERS, BASED ON CITERS 300

COLLABORATION IN AFRICA: NETWORKS OR CLUSTERS? 316

COLLABORATIVE INNOVATIVE NETWORKS: INFLUENCE AND PERFORMANCE 328

COMPARATIVE STUDY ON STRUCTURE AND CORRELATION AMONG BIBLIOMETRICS CO-OCCURRENCE NETWORKS AT AUTHOR-LEVEL 339

COMPARING BOOK CITATIONS IN HUMANITIES JOURNALS TO LIBRARY HOLDINGS: SCHOLARLY USE VERSUS 'PERCEIVED CULTURAL BENEFIT' (RIP) 353

A COMPARISON OF TWO HIGHLY DETAILED, DYNAMIC, GLOBAL MODELS AND MAPS OF SCIENCE 361

A COMPREHENSIVE INDEX TO ASSESS A SINGLE ACADEMIC PAPER IN THE CONTEXT OF CITATION NETWORK (RIP) 377

THE CONSTRUCTION OF THE ACADEMIC WORLD-SYSTEM: REGRESSION AND SOCIAL NETWORK APPROACHES TO ANALYSIS OF INTERNATIONAL ACADEMIC TIES..... 389

CONSTRUCTION OF TYPOLOGY OF SUB-DISCIPLINES BASED ON KNOWLEDGE INTEGRATION 404

CONTRIBUTION AND INFLUENCE OF PROCEEDINGS PAPERS TO CITATION IMPACT IN SEVEN CONFERENCE AND JOURNAL-DRIVEN SUB-FIELDS OF ENERGY RESEARCH 2005-11 (RIP)..... 418

CORE-PERIPHERY STRUCTURES IN NATIONAL HIGHER EDUCATION SYSTEMS. A CROSS-COUNTRY ANALYSIS USING INTERLINKING DATA 426

CORRELATION AMONG THE SCIENTIFIC PRODUCTION, SUPERVISIONS AND PARTICIPATION IN DEFENSE EXAMINATION COMMITTEES IN THE BRAZILIAN PHYSICISTS COMMUNITY (RIP) 447

COUNTING PUBLICATIONS AND CITATIONS: IS MORE ALWAYS BETTER?..... 455

COVERAGE AND ADOPTION OF ALTMETRICS SOURCES IN THE BIBLIOMETRIC COMMUNITY 468

CROWDSOURCING THE NAMES-GAME: A PROTOTYPE FOR NAME DISAMBIGUATION OF AUTHOR-INVENTORS (RIP) 484

DETECTING THE HISTORICAL ROOTS OF RESEARCH FIELDS BY REFERENCE PUBLICATION YEAR SPECTROSCOPY (RPYS) 493

DETECTION OF NEXT RESEARCHES USING TIME TRANSITION IN FLUORESCENT PROTEINS 507

DIFFERENCES AND SIMILARITIES IN USAGE VERSUS CITATION BEHAVIOURS OBSERVED FOR FIVE SUBJECT AREAS 519

DIFFERENCES IN CITATION IMPACT ACROSS COUNTRIES 536

DIRECTIONAL RETURNS TO SCALE OF BIOLOGICAL INSTITUTES IN CHINESE ACADEMY OF SCIENCES..... 551

DISCIPLINARY DIFFERENCES IN TWITTER SCHOLARLY COMMUNICATION..... 567

THE DISCOVERY OF ‘THE UBIQUITIN-MEDIATED PROTEOLYTIC SYSTEM’: AN EXAMPLE OF REVOLUTIONARY SCIENCE? (RIP)..... 583

THE DISTRIBUTION OF REFERENCES IN SCIENTIFIC PAPERS: AN ANALYSIS OF THE IMRAD STRUCTURE..... 591

DO BLOG CITATIONS CORRELATE WITH A HIGHER NUMBER OF FUTURE CITATIONS? (RIP) 604

DO NON-SOURCE ITEMS MAKE A DIFFERENCE IN THE SOCIAL SCIENCES? 612

DOWNLOAD VS. CITATION VS. READERSHIP DATA: THE CASE OF AN INFORMATION SYSTEMS JOURNAL 626

DYNAMICS OF SCIENCE AND TECHNOLOGY CATCH-UP BY SELECTED ASIAN ECONOMIES: A COMPOSITE ANALYSIS COMBINING SCIENTIFIC PUBLICATIONS AND PATENTING DATA 635

THE EFFECT OF BOOMING COUNTRIES ON CHANGES IN THE RELATIVE SPECIALIZATION INDEX (RSI) ON COUNTRY LEVEL ... 654

THE EFFECT OF FUNDING MODES ON THE QUALITY OF KNOWLEDGE PRODUCTION 664

EFFECTS OF RESEARCH FUNDING, GENDER AND TYPE OF POSITION ON RESEARCH COLLABORATION NETWORKS: A MICRO-LEVEL STUDY OF CANCER RESEARCH AT LUND UNIVERSITY 677

EVALUATING KNOWLEDGE PRODUCTION SYSTEMS: MULTIDISCIPLINARITY AND HETEROGENEITY IN HEALTH SCIENCES RESEARCH 690

EVALUATING THE WEB RESEARCH DISSEMINATION OF EU ACADEMICS: A MULTI-DISCIPLINE OUTLINK ANALYSIS OF ONLINE CVS 705

AN EXAMINATION OF THE POSSIBILITIES THAT ALTMETRIC METHODS OFFER IN THE CASE OF THE HUMANITIES (RIP) 720

EXPLORING QUANTITATIVE CHARACTERISTICS OF PATENTABLE APPLICATIONS USING RANDOM FORESTS 728

EXTENDING AUTHOR CO-CITATION ANALYSIS TO USER INTERACTION ANALYSIS: A CASE STUDY ON INSTANT MESSAGING GROUPS 742

EXTENDING CITER-BASED ANALYSIS TO JOURNAL IMPACT EVALUATION 755

FIELD-NORMALIZATION OF IMPACT FACTORS: RESCALING *VERSUS* FRACTIONALLY COUNTED 769

FUNDING ACKNOWLEDGEMENTS FOR THE GERMAN RESEARCH FOUNDATION (DFG). THE DIRTY DATA OF THE WEB OF SCIENCE DATABASE AND HOW TO CLEAN IT UP. 784

GENDER AND ACADEMIC ROLES IN GRADUATE PROGRAMS: ANALYSES OF BRAZILIAN GOVERNMENT DATA 796

GENDER INEQUALITY IN SCIENTIFIC PRODUCTION (RIP) 811

GENETICALLY MODIFIED FOOD RESEARCH IN CHINA: INTERACTIONS BETWEEN AUTHORS FROM SOCIAL SCIENCES AND NATURAL SCIENCES 819

A GLOBAL OVERVIEW OF COMPLEX NETWORKS RESEARCH ACTIVITIES 831

HOW ARE COLLABORATION AND PRODUCTIVITY CORRELATED AT VARIOUS CAREER STAGES OF SCIENTISTS? 847

HOW TO COMBINE TERM CLUMPING AND TECHNOLOGY ROADMAPPING FOR NEWLY EMERGING SCIENCE & TECHNOLOGY COMPETITIVE INTELLIGENCE: THE SEMANTIC TRIZ TOOL AND CASE STUDY 861

HOW WELL DEVELOPED ARE ALTMETRICS? CROSS-DISCIPLINARY ANALYSIS OF THE PRESENCE OF 'ALTERNATIVE METRICS' IN SCIENTIFIC PUBLICATIONS (RIP) 876

INTERMEDIATE-CLASS UNIVERSITY RANKING SYSTEM: APPLICATION TO MAGHREB UNIVERSITIES (RIP) 885

IDENTIFYING EMERGING RESEARCH FIELDS WITH PRACTICAL APPLICATIONS VIA ANALYSIS OF SCIENTIFIC AND TECHNICAL DOCUMENTS 896

IDENTIFYING EMERGING TECHNOLOGIES: AN APPLICATION TO NANOTECHNOLOGY 912

IDENTIFYING EMERGING TOPICS BY COMBINING DIRECT CITATION AND CO-CITATION 928

IDENTIFYING LONGITUDINAL DEVELOPMENT AND EMERGING TOPICS IN WIND ENERGY FIELD 941

THE IMPACT OF CORE DOCUMENTS: A CITATION ANALYSIS OF THE 2003 SCIENCE CITATION INDEX CORE-DOCUMENT POPULATION 955

IMPACT OF META-ANALYTICAL STUDIES, STANDARD ARTICLES AND REVIEWS: SIMILARITIES AND DIFFERENCES 966

THE IMPACT OF R&D ACTIVITIES ON HOSPITAL OUTCOMES (RIP) 978

INDUSTRY RESEARCH PRODUCTION AND LINKAGES WITH ACADEMIA: EVIDENCE FROM UK SCIENCE PARKS 985

INFLUENCE OF UNIVERSITY MERGERS AND THE NORWEGIAN PERFORMANCE INDICATOR ON OVERALL DANISH CITATION IMPACT 2000-12 1003

INFORMATION AND LIBRARY SCIENCE, CHANGES THAT INFLUENCED IT'S NEW CHARACTER, DIRECTION AND RESEARCH: A BIBLIOMETRIC STUDY, 1985-2006 1019

AN INFORMETRIC STUDY OF KNOWLEDGE FLOW AMONG SCIENTIFIC FIELDS (RIP) 1030

INTERACTIVE OVERLAYS OF JOURNALS AND THE MEASUREMENT OF INTERDISCIPLINARITY 1037

INTERDISCIPLINARY RESEARCH AND THE PRODUCTION OF LOCAL KNOWLEDGE: EVIDENCE FROM A DEVELOPING COUNTRY..... 1053

INTERNATIONAL COMPARATIVE STUDY ON NANOFILTRATION MEMBRANE TECHNOLOGY BASED ON RELEVANT PUBLICATIONS AND PATENTS..... 1069

IN-TEXT AUTHOR CITATION ANALYSIS: AN INITIAL TEST (RIP) 1082

KNOWLEDGE CAPTURE MECHANISMS IN BIOVENTURE CORPORATIONS: A CASE STUDY 1090

LEAD-LAG TOPIC EVOLUTION ANALYSIS: PREPRINTS VS. PAPERS (RIP) 1106

LITERATURE RETRIEVAL BASED ON CITATION CONTEXT..... 1114

MAPPING THE EVOLVING PATTERNS OF PATENT ASSIGNEES' COLLABORATION NETWORK AND IDENTIFYING THE COLLABORATION POTENTIAL 1135

MATCHING BIBLIOGRAPHIC DATA FROM PUBLICATION LISTS WITH LARGE DATABASES USING N-GRAMS (RIP) 1151

MATHEMATICAL CHARACTERIZATIONS OF THE WU- AND HIRSCH-INDICES USING TWO TYPES OF MINIMAL INCREMENTS 1159

MEASURING INTERNATIONALISATION OF BOOK PUBLISHING IN THE SOCIAL SCIENCES AND HUMANITIES USING THE BARYCENTRE METHOD (RIP) 1170

MEASURING THE ACADEMIC IMPACT OF RESEARCHERS BY COMBINED CITATION AND COLLABORATION IMPACT 1177

MEASURING THE EXTENT TO WHICH A RESEARCH DOMAIN IS SELF-CONTAINED 1188

A METHOD FOR TEXT NETWORK ANALYSIS: TESTING, DEVELOPMENT AND APPLICATION TO THE INVESTIGATION OF PATENT PORTFOLIOS (RIP) 1202

MISFITS? RESEARCH CLASSIFICATION IN RESEARCH EVALUATION: VISUALIZING JOURNAL CONTENT WITHIN FIELDS OF RESEARCH CODES..... 1210

MODEL TO SUPPORT THE INFORMATION RETRIEVAL PROCESS OF THE SCIENTIFIC PRODUCTION AT DEPARTMENTAL-LEVEL OR FACULTY-LEVEL OF UNIVERSITIES 1225

MOST BORROWED IS MOST CITED? LIBRARY LOAN STATISTICS AS A PROXY FOR MONOGRAPH SELECTION IN CITATION INDEXES (RIP) 1237

MOTIVATION FOR HYPERLINK CREATION USING INTER-PAGE RELATIONSHIPS 1253

MOVING FROM PERIPHERY TO CORE IN SCIENTIFIC NETWORKS: EVIDENCE FROM EUROPEAN INTER-REGIONAL COLLABORATIONS, 1999-2007 (RIP) 1270

NANO-ENHANCED DRUG DELIVERY (NEDD) RESEARCH PATTERN FOR TWO LEADING COUNTRIES: US AND CHINA 1278

NANOTECHNOLOGY AS GENERAL PURPOSE TECHNOLOGY 1291

NEVIEWER: A NEW SOFTWARE FOR ANALYZING THE EVOLUTION OF RESEARCH TOPICS 1307

THE NUANCED NATURE OF E-PRINT USE: A CASE STUDY OF ARXIV 1321

ON THE DETERMINANTS OF RESEARCH PERFORMANCE: EVIDENCE FROM ECONOMIC DEPARTMENTS OF FOUR EUROPEAN COUNTRIES (RIP) 1334

OPEN DATA AND OPEN CODE FOR BIG SCIENCE OF SCIENCE STUDIES 1342

OPTIMIZING RESEARCH IMPACT BY ALLOCATING FUNDING TO RESEARCHER GRANT PORTFOLIOS: SOME EVIDENCE ON A POLICY OPTION (RIP) 1357

PATENTS IN NANOTECHNOLOGY: AN ANALYSIS USING MACRO-INDICATORS AND FORECASTING CURVES 1363

THE PATTERNS OF INDUSTRY-UNIVERSITY-GOVERNMENT COLLABORATION IN PHOTOVOLTAIC TECHNOLOGY 1379

PERFORMING INFORMETRIC ANALYSIS ON INFORMATION RETRIEVAL TEST COLLECTIONS: PRELIMINARY EXPERIMENTS IN THE PHYSICS DOMAIN (RIP) 1392

POSSIBILITIES OF FUNDING ACKNOWLEDGEMENT ANALYSIS FOR THE BIBLIOMETRIC STUDY OF RESEARCH FUNDING ORGANIZATIONS: CASE STUDY OF THE *AUSTRIAN SCIENCE FUND (FWF)* 1401

PREDICTING AND RECOMMENDING POTENTIAL RESEARCH COLLABORATIONS 1409

PUBLICATION BIAS IN MEDICAL RESEARCH: ISSUES AND COMMUNITIES 1419

QUANTITATIVE EVALUATION OF ALTERNATIVE FIELD NORMALIZATION PROCEDURES 1431

A RELATION BETWEEN POWER LAW DISTRIBUTIONS AND HEAPS' LAW 1445

THE RELATIONSHIP BETWEEN COLLABORATION AND PRODUCTIVITY FOR LONG-TERM INFORMATION SCIENCE RESEARCHERS (RIP) 1461

RELATIONSHIP BETWEEN DOWNLOADS AND CITATION AND THE INFLUENCE OF LANGUAGE 1469

RELEVANCE AND FOCUS SHIFT: NEW METRICS FOR THE GRANT EVALUATION PROCESS PILOT TESTED ON NIH GRANT APPLICATIONS (RIP) 1485

RELEVANCE DISTRIBUTIONS ACROSS BRADFORD ZONES: CAN BRADFORDIZING IMPROVE SEARCH? 1493

RESEARCH COLLABORATION AND PRODUCTION OF EXCELLENCE: FINLAND 1995-2009 1506

RESEARCH PERFORMANCE ASSESSMENT USING NORMALIZATION METHOD BASED ON SCI DATABASE (RIP) 1528

RETHINKING RESEARCH EVALUATION INDICATORS AND METHODS FROM AN ECONOMIC PERSPECTIVE: THE FSS INDICATOR AS A PROXY OF PRODUCTIVITY 1536

THE ROLE OF NATIONAL UNIVERSITY RANKINGS IN AN INTERNATIONAL CONTEXT: THE CASE OF THE I-UGR RANKINGS OF SPANISH UNIVERSITIES 1550

SCIENCE DYNAMICS: NORMALIZED GROWTH CURVES, SHARPE RATIOS, AND SCALING EXPONENTS 1566

SCIENTIFIC POLICY IN BRAZIL: EXPLORATORY ANALYSIS OF ASSESSMENT CRITERIA (RIP) 1578

'SEED+EXPAND': A VALIDATED METHODOLOGY FOR CREATING HIGH QUALITY PUBLICATION OEUVRES OF INDIVIDUAL RESEARCHERS 1587

THE SHORTFALL IN COVERAGE OF COUNTRIES' PAPERS IN THE SOCIAL SCIENCES CITATION INDEX COMPARED WITH THE SCIENCE CITATION INDEX 1601

SOCIAL DYNAMICS OF RESEARCH COLLABORATION: NORMS, PRACTICES, AND ETHICAL ISSUES IN DETERMINING CO-AUTHORSHIP RIGHTS (RIP) 1613

SOFTWARE PATENTING IN ASIA 1622

SUPPLY AND DEMAND IN SCHOLARLY PUBLISHING: AN ANALYSIS OF FACTORS ASSOCIATED WITH JOURNAL ACCEPTANCE RATES (RIP) 1640

A SYSTEMATIC EMPIRICAL COMPARISON OF DIFFERENT APPROACHES FOR NORMALIZING CITATION IMPACT INDICATORS 1649

THE TIPPING POINT – OPEN ACCESS COMES OF AGE 1665

TO WHAT EXTENT CAN RESEARCHERS' INTERNATIONAL MOVEMENT BE GRASPED FROM PUBLISHED DATA SOURCES? . 1681

TO WHAT EXTENT IS THE H-INDEX INCONSISTENT? IS STRICT CONSISTENCY A REASONABLE REQUIREMENT FOR A SCIENTOMETRIC INDICATOR? 1696

TOWARD A TIME-SENSITIVE MESOSCOPIC ANALYSIS OF CO-AUTHOR NETWORKS: A CASE STUDY OF TWO RESEARCH SPECIALTIES 1711

TOWARDS THE DEVELOPMENT OF AN INDICATOR OF CONFORMITY 1726

TRACING RESEARCH PATHS OF SCIENTISTS BY MEANS OF CITATIONS 1738

TRACKING ACADEMIC REGIONAL WORKFORCE RETENTION THROUGH AUTHOR AFFILIATION DATA 1746

TRENDS OF INTELLECTUAL AND COGNITIVE STRUCTURES OF STEM CELL RESEARCH: A STUDY OF BRAZILIAN SCIENTIFIC PUBLICATIONS 1759

USE OF ELECTRONIC JOURNALS IN UNIVERSITY LIBRARIES: AN ANALYSIS OF OBSOLESCENCE REGARDING CITATIONS AND ACCESS 1772

USING MONTE CARLO SIMULATIONS TO ASSESS THE IMPACT OF AUTHOR NAME DISAMBIGUATION QUALITY ON DIFFERENT BIBLIOMETRIC ANALYSES. 1784

VISUALIZING AND COMPARING THE DEVELOPMENT OF SCIENTIFIC INSTRUMENTATION VS ENGINEERING INSTRUMENTATION 1792

WEB BASED IMPACT MEASURES FOR INSTITUTIONAL REPOSITORIES 1806

WHAT IS THE IMPACT OF SCALE AND SPECIALIZATION ON THE RESEARCH EFFICIENCY OF EUROPEAN UNIVERSITIES? 1817

WHICH FACTORS HELP TO PRODUCE HIGH IMPACT RESEARCH? A COMBINED STATISTICAL MODELLING APPROACH 1830

POSTERS 1845

THE 2-YEAR MAXIMUM JOURNAL IMPACT FACTOR 1847

ACCURACY ASSESSMENT FOR BIBLIOGRAPHIC DATA 1850

ANALYSIS OF SEARCH RESULTS FOR THE CLARIFICATION AND IDENTIFICATION OF TECHNOLOGY EMERGENCE (AR-CITE) 1854

APPLICATIONS AND RESEARCHES OF GIS TECHNOLOGIES IN BIBLIOMETRICS 1857

APPROPRIATE COVERAGE OF SCHOLARLY PUBLISHING IN THE SOCIAL SCIENCES AND HUMANITIES - A EUROPEAN OVERVIEW 1861

ARE REGISTERED AUTHORS MORE PRODUCTIVE? 1864

ARE THE BRIC AND MITS COUNTRIES IMPROVING THEIR PRESENCE IN THE INTERNATIONAL SCIENCE? 1868

JOURNAL IMPACT FACTOR, EIGENFACTOR, JOURNAL INFLUENCE AND ARTICLE INFLUENCE 1871

ASEP ANALYTICS. A SOURCE FOR EVALUATION AT THE ACADEMY OF SCIENCES OF THE CR 1874

ASSESSING AN INTERVAL OF CONFIDENCE TO COMPILE TIME-DEPENDENT PATENT INDICATORS IN NANOTECHNOLOGY 1877

BIBLIOMETRIC INDICATORS OF YOUNG AUTHORS IN ASTROPHYSICS: CAN LATER STARS BE PREDICTED? 1881

BIOLOGICAL SCIENCES PRODUCTION: A COMPARATIVE STUDY ON THE MODALITIES OF FULL PHD IN BRAZIL OR ABROAD 1884

A CITATION ANALYSIS ON MONOGRAPHS IN THE FIELD OF SCIENTOMETRICS, INFORMETRICS AND BIBLIOMETRICS IN CHINA (1987-2010) 1887

CITATION PATTERNS FOR SOCIAL SCIENCES AND HUMANITIES PUBLICATIONS 1891

COLLABORATION IN THE SOCIAL SCIENCES AND HUMANITIES: EDITED BOOKS IN ECONOMICS, HISTORY AND LINGUISTICS 1894

THE COLLECTIVE CONSEQUENCES OF SCIENTIFIC FRAUD: AN ANALYSIS OF BIOMEDICAL RESEARCH 1897

COMPARING NATIONAL DISCIPLINARY STRUCTURES: A QUANTITATIVE APPROACH 1900

COMPREHENSIVENESS AND ACCURACY OF DOCUMENT TYPES: COMPARISON IN WEB OF SCIENCE AND SCOPUS AGAINST PUBLISHER'S DEFINITION 1905

CONTRIBUTION OF BRAZILIAN SCIENTIFIC PRODUCTION TO MAINSTREAM SCIENCE IN THE FIELD OF MATHEMATICS: A SCIENTOMETRICS ANALYSIS (2002-2011) 1908

CO-OCCURRENCE BETWEEN AUTHORS' AFFILIATION AND JOURNAL: ANALYSIS BASED ON 2-MODE NETWORK 1912

COST ANALYSIS OF E –JOURNALS, BASED ON THE SCIENTIFIC COMMUNITIES USAGE OF SCIENCE DIRECT ONLINE DATABASE WITH SPECIAL REFERENCE TO BANARAS HINDU UNIVERSITY LIBRARY, INDIA 1915

A COVERAGE OVERLAP STUDY ON CITATION INDEX: COMMERCIAL DATABASES AND OPEN ACCESS SYSTEMS 1918

FACTORS RELATED TO GENDER DIFFERENCES IN SCIENCE: A CO-WORD ANALYSIS 1922

THE CROSSCHECK PLAGIARISM SYSTEM: A BRIEF STUDY FOR SIMILARITY 1925

CUMULATIVE CAPABILITIES IN COLOMBIAN UNIVERSITIES: AN EVALUATION USING SCIENTIFIC PRODUCTIVITY 1928

A DESCRIPTIVE STUDY OF INACCURACY IN ARTICLE TITLES ON BIBLIOMETRICS PUBLISHED IN BIOMEDICAL JOURNALS 1932

DIFFUSION OF BRAZILIAN STATISTIC INFORMATION 1935

DISCOVERING AUTHOR IMPACT: A NOVEL INDICATOR BASED ON CITATION IDENTITY 1938

DO NEW SCIENTISTS PREFER COLLABORATING WITH OLD SCIENTISTS? AND VICE VERSA? 1941

DO SMALL AND MEDIUM SIZED BUSINESSES CLAIM FOR SMALL ENTITY STATUS? THE CASE OF MIT AND STANFORD UNIVERSITY SPINOFFS 1944

DOES SCIENTIFIC KNOWLEDGE PLAY A ROLE IN PUBLIC POLICIES? A CONTRIBUTION OF SCIENTOMETRICS TO POLITICAL SCIENCE: THE CASE OF HTA. 1947

THE EARLIEST PRIORITY SELECTOR FOR COMPILING PATENT INDICATORS..... 1950

EFFICIENCIES IN NATIONAL SCIENTIFIC PRODUCTIVITY WITH RESPECT TO MANPOWER AND FUNDING IN SCIENCE..... 1954

EMERGENCE OF KEYWORDS IN WEB OF SCIENCE VS. WIKIPEDIA 1957

ENTROPY-BASED DISCIPLINARITY INDICATOR: ROLE TAXONOMY OF JOURNALS IN SCIENTIFIC COMMUNICATION SYSTEMS..... 1960

THE EPIDEMIC OF RENAL DISEASE –AN EVALUATION OF STATUS (2005-2009)..... 1963

EUROPEAN HIGHLY CITED SCIENTISTS' PRESENCE IN THE SOCIAL WEB..... 1966

EVALUATING THE INVENTIVE ACTIVITY OF FOREIGN R&D CENTERS IN ISRAEL: LINKING PATSTAT TO FIRM LEVEL DATA 1970

EVALUATION OF RESEARCH IN SPAIN: BIBLIOMETRIC INDICATORS USED BY MAJOR SPANISH RESEARCH ASSESSMENT AGENCIES 1973

AN EXPERIENCE OF THE INCLUSION A NEW METHODOLOGY IN SELECTING THE REVIEWERS FOR GRANT APPLICATIONS..... 1976

EXPLORING INTERDISCIPLINARITY IN ECONOMICS THROUGH ACADEMIC GENEALOGY: AN EXPLORATORY STUDY 1979

FEATURES OF INDEX TERMS AND NATURAL LANGUAGE WORDS FROM THE PERSPECTIVE OF EXTRACTED TOPICS 1983

FROM CATEGORICAL TO RELATIONAL DIVERSITY – EXPLORING NEW APPROACHES TO MEASURING SCIENTIFIC DIVERSITY 1986

FULLERENE AND COLD FUSION: BIBLIOMETRIC DISCRIMINATION BETWEEN NORMAL AND PATHOLOGICAL SCIENCE 1989

GEOGRAPHICAL ORIENTATION AND IMPACT OF FINLAND'S INTERNATIONAL CO-PUBLICATIONS..... 1992

GLOBAL RESEARCH STATUS IN LEADING NUCLEAR SCIENCE AND TECHNOLOGY JOURNALS DURING 2001–2010: A BIBLIOMETRIC ANALYSIS BASED ON ISI WEB OF SCIENCE..... 1995

GROUPS OF HIGHLY CITED PUBLICATIONS: STABILITY IN CONTENT WITH CITATION WINDOW LENGTH 1998

HEAPS' LAW: A DYNAMIC PERSPECTIVE FROM SIMON'S MODEL 2001

HOW EFFECTIVE IS THE KNOWLEDGE TRANSFER OF A PUBLIC RESEARCH ORGANIZATION (PRO)? FIRST EMPIRICAL EVIDENCE FROM THE SPANISH NATIONAL RESEARCH COUNCIL..... 2004

HOW MUCH MATHEMATICS IS IN THE *BIG TWO* AND WHERE IS IT LOCATED? 2008

IDENTIFICATION METHOD ON LOW QUALITY PATENTS AND APPLICATION IN CHINA 2011

IMPACT AND VISIBILITY OF SA'S RESEARCH JOURNALS: ASSESSING THE 2008 EXPANSION IN COVERAGE OF THE THOMSON REUTERS DATABASES 2014

IMPACT OF BRAIN DRAIN ON SCIENCE PRODUCTION: A CASE STUDY OF IRANIAN EDUCATED MIGRANTS IN THE CONTEXT OF SCIENCE PRODUCTION IN CANADA 2017

AN INDEX TO QUALIFY HUMAN RESOURCES OF AN ENTERPRISES CLUSTER 2020

AN INTERPRETABLE AXIOMATIZATION OF THE HIRSCH-INDEX2024

INTERPRETING EPISTEMIC AND SOCIAL CULTURAL IDENTITIES OF DISCIPLINES WITH MACHINE LEARNING MODELS OF METADISCOURSE 2027

AN INVESTIGATION OF SCIENTIFIC COLLABORATION BETWEEN IRAN AND OTHER MENA COUNTRIES AND ITS RELATIONSHIP WITH ECONOMIC INDICATORS 2031

KEYWORD-QUERY EXPANSION USING CITATION CLUSTERS FOR PAPER INFORMATION RETRIEVAL 2034

KNOWLEDGE COMBINATION FORECASTING BETWEEN DIFFERENT TECHNOLOGICAL FIELDS 2037

LANGUAGE PREFERENCE IN SOCIOLOGICAL RESEARCH PUBLISHED BY VARIOUS EUROPEAN NATIONALITIES 2040

LEADERS AND PARTNERS IN INTERNATIONAL COLLABORATION AND THEIR INFLUENCE ON RESEARCH IMPACT 2044

MEASURING INTERDISCIPLINARITY OF RESEARCH GRANT APPLICATIONS. AN INDICATOR DEVELOPED TO MODEL THIS SELECTION CRITERION IN THE ERC'S PEER-REVIEW PROCESS .. 2048

MEASURING THE QUALITY OF ACADEMIC MENTORING 2051

A MODEL BASED ON BIBLIOMETRIC INDICATORS: THE PREDICTIVE POWER 2054

MONITORING OF INDIAN RESEARCH PAPERS: ON THE BASIS OF MAJOR GLOBAL SECONDARY SERVICES 2057

NANOSCIENCE AND NANOTECHNOLOGY IN SCOPUS: JOURNAL IDENTIFICATION AND VISUALIZATION 2061

A NEW APPROACH FOR AUTOMATED AUTHOR DISCIPLINE CATEGORIZATION AND EVALUATION OF CROSS-DISCIPLINARY COLLABORATIONS FOR GRANT PROGRAMS 2066

NORMALIZED INDICATORS OF THE INTERNATIONAL BRAZILIAN RESEARCH: A SCIENTOMETRIC STUDY OF THE PERIOD BETWEEN 1996 AND 2011 2069

ON THE DEFINITION OF A REVIEW, AND DOES IT MATTER? 2072

AN ONLINE SYSTEM FOR MANAGEMENT AND MONITORING OF EXTRAMURAL PROPOSALS FOR FUNDING BY ICMR – A CASE STUDY 2075

PAPERS PUBLISHED IN PNAS REFLECT THE HIERARCHY OF THE SCIENCES 2080

A RESEARCH PROFILE FOR A PROMISING EMERGING INDUSTRY – NANO-ENABLED DRUG DELIVERY 2083

THE P-INDEX: HIRSCH INDEX OF INDIVIDUAL PUBLICATIONS .. 2086

PRELIMINARY ANALYSIS OF THE FINANCIAL ASSISTANCE TO NON-ICMR BIOMEDICAL SCIENTISTS BY INDIAN COUNCIL OF MEDICAL RESEARCH (ICMR) 2089

THE PRODUCTIVITY AND IMPACT OF ASTRONOMICAL TELESCOPES – A BIBLIOMETRIC STUDY FOR 2007 – 2011 2092

PROFILES OF PRODUCTION, IMPACT, VISIBILITY AND COLLABORATION OF THE SPANISH UNIVERSITY SYSTEM IN SOCIAL SCIENCES AND HUMANITIES 2095

PROTOTYPICAL STRATEGY FOR HIGH-LEVEL CITATION-ANALYSES: A CASE STUDY ON THE RECEPTION OF ENGLISH-LANGUAGE JOURNAL ARTICLES FROM PSYCHOLOGY IN THE GERMAN-SPEAKING COUNTRIES 2099

A QUANTITATIVE ANALYSIS OF ANTARCTIC RELATED ARTICLES IN HUMANITIES AND SOCIAL SCIENCES APPEARING IN THE WORLD CORE JOURNALS 2102

THE RELATIONSHIP BETWEEN A TOPIC'S INTERDISCIPLINARITY AND ITS INNOVATIVENESS 2105

HIERARCHICAL CLUSTERING PHRASED IN GRAPH THEORY: MINIMUM SPANNING TREES, REGIONS OF INFLUENCE, AND DIRECTED TREES. 2109

RESEARCH SECTORS INVOLVED IN CUBAN SCIENTIFIC OUTPUT 2003-2007 2113

RESEARCH TRENDS IN GENETICS: SCIENTOMETRIC PROFILE OF SELECTED ASIAN COUNTRIES 2117

THE RISE AND FALL OF GREECE'S RESEARCH PUBLICATION RECORD: THE LAST 30 YEARS 2120

THE ROLE OF COGNITIVE DISTINCTIVENESS ON CO-AUTHOR SELECTION AND THE INFLUENCE OF CO-AUTHORING ON COGNITIVE STRUCTURE: A MULTI-AGENT SIMULATION APPROACH 2124

SCIENTIFIC PRODUCTION AND INTERNATIONAL COLLABORATION ON SOLAR ENERGY IN SPAIN AND GERMANY (1995-2009) 2126

SCIENTIFIC PRODUCTION OF TOP BRAZILIAN RESEARCHERS IN BIOCHEMISTRY, PHYSIOLOGY, PHARMACOLOGY AND BIOPHYSICS 2129

A SIMPLE METHOD TO ASSESS THE QUALITY OF ANY UNIFICATION PROCESS 2132

STRUCTURE ANALYSIS OF SMALL PATENT CITATION NETWORK AND MAPPING TECHNOLOGICAL TRAJECTORIES 2136

STRUCTURE OF INTERDISCIPLINARY RESEARCH: COMPARING LM AND LDA 2140

THE STUDY AND ASSESSMENT OF RESEARCH PERFORMANCE AT THE MICRO LEVEL: THE AGE PHASE DYNAMICS APPROACH 2143

THE SUBJECT CATEGORIES NORMALIZED IMPACT FACTOR 2146

SUCCESS DETERMINANTS OF FULL-TIME RESEARCHERS AT HOSPITALS. A PERCEPTIONS-BASED STUDY 2149

SURFING THE SEMANTIC WEB 2152

TEMPORAL EVOLUTION, STRUCTURAL FEATURES AND IMPACT OF STANDARD ARTICLES AND PROCEEDINGS PAPERS. A CASE STUDY IN BLENDED LEARNING. 2156

TESTING COMPOSITE INDICATORS FOR THE SCIMAGO INSTITUTIONS RANKING 2159

A TEXT MINING APPROACH EXPLORING ACKNOWLEDGEMENTS OF PAPERS 2162

REGULARITY IN THE TIME-DEPENDENT DISTRIBUTION OF THE PERCENTAGE OF UNCITED ARTICLES: AN EMPIRICAL PILOT STUDY BASED ON THE SIX JOURNALS 2165

TOPOLOGICAL TOPIC TRACKING – A COMPARATIVE ANALYSIS 2168

TOWARDS AN AUTHOR-TOPIC-TERM-MODEL VISUALIZATION OF 100 YEARS OF GERMAN SOCIOLOGICAL SOCIETY PROCEEDINGS 2171

USE FREQUENCIES OF NOMINALIZATIONS IN SCIENTIFIC WRITING IN BRAZILIAN PORTUGUESE LANGUAGE AS POLITENESS STRATEGIES AND THEIR INDEX ROLE IN THE SUBJECT INDEXING 2174

A VISUALIZATION TOOL FOR TOPIC EVOLUTION AMONG RESEARCH FIELDS 2178

VISUALIZING THE RESEARCH DOMAIN ON SCIENTOMETRICS (1978-2012) 2182

WEB 2.0 TOOLS FOR NETWORK MANAGEMENT AND PATENT ANALYSIS FOR HEALTH PUBLIC 2185

WEIGHTING CO-CITATION PROXIMITY BASED ON CITATION CONTEXT 2189

WHAT MEANS, IN NUMBERS, A GOLD STANDARD BIOCHEMISTRY DEPARTMENT TO NATIONAL AGENCIES OF RESEARCH FOMENTATION IN BRAZIL?..... 2193

WHEN INNOVATION INDICATORS MEET SPIN-OFF COMPANIES: A BRIEF REVIEW AND IMPROVEMENT PROPOSAL..... 2196

WHERE NATURAL SCIENCES (PHYSICS) MADE IN THE WORLD AND IN RUSSIA: 3-DECADES DYNAMICS 2200

AUTHOR INDEX 2203

MAPPING THE EVOLVING PATTERNS OF PATENT ASSIGNEES' COLLABORATION NETWORK AND IDENTIFYING THE COLLABORATION POTENTIAL

Yunwei Chen[1,2*], Shu Fang[1*]

[1]*chenyw@clas.ac.cn; fangsh@clas.ac.cn
1 Chengdu Library of the Chinese Academy of Sciences, Chengdu, 610041 (China)
2 University of Chinese Academy of Sciences, Beijing, 100049 (China)

Abstract
The purpose of this article is to map the evolving patterns of patent assignees' collaboration networks and build a Latent Collaboration Index (LCI) model for evaluating the collaboration probability among patent assignees. The demonstration process was carried on the patents of field of industrial biotechnology (IB) from 2000 to 2010. The studies deployed two different network analysis tools, NWB (NWB Team, 2006) and Thomson Data Analyzer (TDA), and used ISI Derwent Innovations IndexSM (DII) to collect all the patents of IB. The results shows that the assignees in the field of IB grew steadily while the number of patents had decreased slowly year by year after it reached peak in 2002 and 2003. Densification and growth analysis, average degree, density and components analysis have showed that the collaboration networks tended to density. Especially from the diameter analysis we might also conclude that the IB field had come into a mature mode after finishing the topological transition occurred in about 2002 or 2003. The nodes of final network had degrees k followed a power law distribution, which implies a preferential linking feature of the network evolving and thus provided a foundation for link prediction from the aspect of network evolving. Basing on this, two network-related parameters had been brought into the LCI model, which were degree and network distance. The values of which are positive and negative for link generation respectively. In addition, types of assignees, geographical distance, topic similarity had also been added into the LCI model. Different type of assignees had different probability to be linked, such as corporation had been collaborated more frequently, universities ranked lowest. Assignees from the same country seemed to be likely to collaborate. It have to been noted that the LCI model is flexible that could be adjusted of the factors or the weights of them according to different subjects, time or data. For instance, the topic similarity factor between assignees would be removed from the LCI model for link prediction in the field of IB because of the poor inference from topic similarity to collaboration.

Conference Topic
Collaboration Studies and Network Analysis (Topic 6)

Introduction

Patents had been studied for a long time and were regarded as providing valuable data for studies of technology progress, innovative activity (Ernst, 2003), identifying technology trends (Segev & Kantola, 2012), strategies for countries or companies (Han & Park, 2006), innovation management and policies designing (Lee, 2010). Patent documents have also been applied to understand the linkages between industries, nations, or technologies in terms of technological innovations and knowledge flow (Lee, 2010).

During these kinds of works many methods had been used including the method of Social Network Analysis (SNA), which just had begun to invade the field of patent analysis (Sternitzke, Bartkowski & Schramm, 2008). Basing on patent information, network nodes could represent inventors, patent assignees or patent filing documents, and so on. For example, cooperation networks between inventors and applicants and citation networks of patent families and applicants had been illustrated by using the method of SNA by Sternitzke, Bartkowski & Schramm (2008). Inter-industrial knowledge flows had been studied by the patent network analysis (Han & Park, 2006).

When the SNA methods were used for patents analysis, most researches had focused on citations networks, collaboration networks and theme maps. For instance, Wartburg, Teichert & Rost (2005) discussed with a methodological reflection and application of multi-stage patent citation analysis for the measurement of inventive progress. Gress (2009) mapped the patent citation network based the USPTO data to discuss the information flow directions. Dou & Bai (2007) analysed the collaboration networks from countries level and applicant level to find the competition relations and developing strategies. Huang & Wang (2010) had examined the small world phenomenon in the patent citation network by a case study of the radio frequency identification network. Other patent mapping work also included Yang, Akers, Yang, *et al.* (2010) who made a patent landscape analysis with visualization output by illustrating two case studies: technology assessment and company assessment.

So far, there was few works focusing on mapping the evolution patterns of collaboration networks of the patent assignees. A correlative work was carried by Hanaki, Nakajima & Ogura (2010), which provided an empirical analysis of evolving networks of successful R& D collaborations in the IT industry. The collaboration links in their work was identified between two companies if there was at least one common inventor listed in the patents owned by the companies. First they showed that the R& D network has become more extensive, more clustered, and more unequal in the sense that 'stars' have emerged in the network. Second, they analysed the effect of the existing network structure in the process of new R&D collaboration formation. Another related work about link prediction was from Guns (2011). He discussed the function of bipartite for link prediction and found that some bipartite predictors form a considerable improvement to their unipartite counterparts.

Therefore, this paper utilizes the present techniques to carry out a pilot study that we undertook to map the evolution patterns of the patents assignees collaboration networks. Furthermore, based on the evolution patterns, this paper also tries to build a model for evaluating the collaboration probability among patent assignees. The patents of Industrial Biotechnology from 2000 to 2010 had been used for demonstration. The collaboration links in our work are identified between two companies if their names occurred at least one time in the patents owned by the assignees.

This paper is organized as follows. Section I is an introduction. Section II introduces the data source & methods. Section III discusses the empirical results of the evolving patterns of patent assignees' collaboration network. Section IV discusses the collaboration potential model strategy and raised a Latent Collaboration Index (LCI) for link prediction. Section V evaluates our findings and presents our conclusions.

Data Source & Methods

We used ISI Derwent Innovations Index^{SM} (DII) to collect all the patents in the field of Industrial Biotechnology from 2000 to 2010 according to the patents definition of Linton, Stone & Wise (2008). Since the publication date of a patent publication is generally 18 months later than the application date, in order to make exact steady data, we used the basic patent year (defined by DII according to the year of the patent family member that had been collected by DII the earliest) to divide the patent data year by year. The studies deployed two different network analysis tools, NWB (NWB Team, 2006) and Thomson Data Analyzer (TDA).

The patents assignees had been cleaned by the following two steps: First, an assignee usually has more than one English name because of the different translation from any other language to English, including different time, different translation organizations or just spell difference. Therefore all the assignees fields have to be cleaned to integrate the different names of one institute to only one unique name. Second, Inventors sometimes would be added to the assignees list at some primary stages of patent applications according to the patent laws of different countries. In order to avoiding the interference of the wrong recording information, this paper deleted all the individual assignees from the assignees list.

The assignees that had 50 or more patents were filtered, which contain 450 core assignees, 5479 collaborators of the core assignees and 78209 patents (58% of all patents from 2000 to 2010). These data were the *basic data* for our analysis.

Furthermore, because this paper analyses 11 years evolving networks, many of the 5479 collaborators of the 450 core assignees have few patents were not appropriate for evolving analysis, therefore, we selected collaborators only had more than 5 patents (588 collaborators) together with the 450 core assignees as *core data* (1038 nodes) for network patterns analysis.

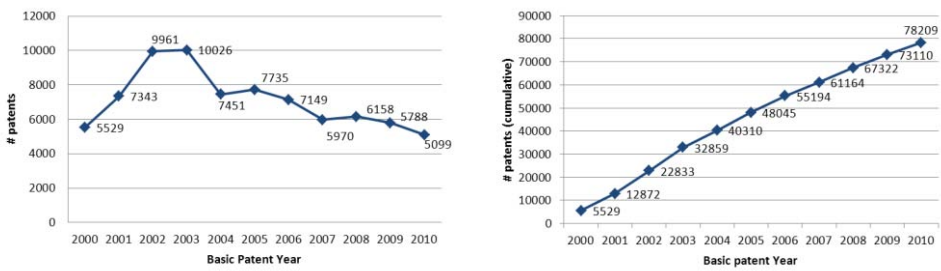
Methodologies deployed in this paper were static in nature, but represent snapshots at certain point of the dynamic investigations, which were used to mapping the patterns of patent collaboration network evolution. The *basic data*

were used in the first two sub-sections of section III. The *core data* were used by the other sections.

Empirical results of the evolving patterns of patent assignees' collaboration network

Mapping the Growth of Industrial Biotechnology (IB) from Patents

The objectives of this section were to collect the global patent data (*basic data*) of IB from 2000 to 2010, to give a general view of the development of IB at the level of applications numbers, annual growth to provide a better understanding of the patent activities in the subject of IB these years.



(a) (b)

Figure 1. Growth of the patents of IB: 2000-2010, annual (a) and cumulative (b)

Figure 1 (a) shows the annual growth of the patents of IB from 2000-2010. From the figure we could see that the patent numbers reached the peak in the year of 2003, and then decreased slowly year by year.

The cumulative number of patents of IB was next investigated and illustrated in Figure 1 (b), which shows approximately a linear growth.

Growth of Assignees

It was to be expected that growth in the number of assignees could show the pattern as the number of patents applications. One hypothesis is that the assignees would increase steadily along with the development of the number of patents applications. However, there were difference in the increasing pattern between number of patents and number of collaborators. The collaborators could recurrence after it first time occurrence, thus in evolution network we did not account it repeatedly. Therefore, we had to count the cumulative number of collaborators from 2000 to 2010. For instance, the collaborators number from 2000-2001 and 2000-2010. The collaborators that had already occurred in earlier years would not be counted again when it reoccurred in later years.

The annual and cumulative number of the assignees (basic data with more than 50 patents and their collaborators) from 2000 to 2010 had been showed in Figure 2. The results show that both the annual and cumulative data of assignees had linear

Growth Laws. As is vividly betrayed in figure 2 of the cumulative curve, the assignees' number increased rapidly with a compound growth rate of 17.3%. It indicated that there were a lot of new assignees emerged in the field of IB.

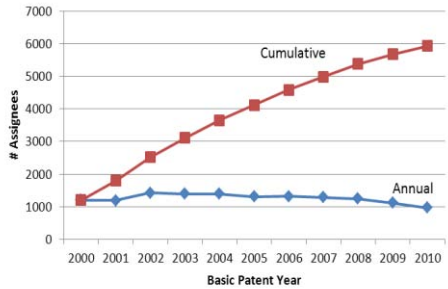


Figure 2. Numbers of the assignees (>50 patents) and collaborators:2000-2010

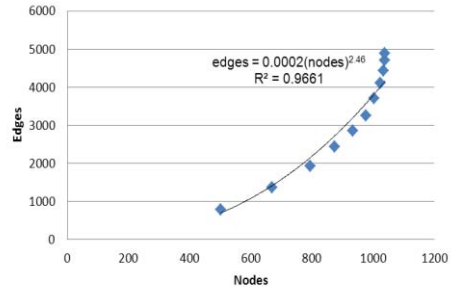


Figure 3. Densification of assignees collaboration networks of IB

Densification and growth

As Bettencourt Kaiser and Kaur (2009) pointed that when fields grew, their networks of collaboration also become denser. This means that the average number of edges per node tends to increase over time. The relation between numbers of nodes and edges has the following simple scaling law with the scaling exponent (α>1) (Bettencourt et al, 2009):

$$\text{edges} = A(\text{nodes})^{\alpha}, \quad (1)$$

A and α are constants. The scaling exponent, α, expresses the densification effect in a way that was independent of scale (number of nodes).

In our work, it is clearly showed in figure 2 that as time gone on, the number of assignees grew. Whether in collaborations networks could also show the feature of scaling exponent (α>1)? For answer this question, we analysed the relation between number of nodes and edges in the IB patent collaboration network using the core data, where nodes represented assignees with more than 50 patents and its collaborators with more than 5 patents, the data in the curve of Fig. 3 started from 2000 and ended at 2010.

We found that the scaling exponent α=2.46, it suggested that the number of ties between assignees grew faster than that of assignees.

Dynamics of the Patents Assignees Collaboration networks

We study the following network statistics: average degree, density, #components, *et al*.

Let N(i) is the set of assignees collaborating with assignee i. The total number of collaboration assignees with assignee i is the degree of assignee i and is defined as $\eta(i) = |N(i)|$. The average degree of a network G is defined by $\eta(G)=\Sigma_{i\in N}\,\eta(i)/n$. The density of a network is defined as the number of links divided by the number of edges in a complete graph with the same number of nodes. For a network G with N nodes, the density D is defined as:

$$D=\frac{2*[\#L(G)]}{N(N-1)} \quad (2)$$

Where #L(G) represented the number of links of the graph G.

For IB assignees collaboration network, the average degree, density and #components show that the network became denser and denser, at the same time some individual components had also merged into one bigger component. The average degree of each assignee had also increased year by year, which meant that the assignees had more and more collaborators.

Building on work by Leskovec, Kleinberg, and Faloutsos (2005) which found that as networks grow and more nodes and edges are added, their effective diameter (as measured by shortest-path length—i.e., the 90th percentile) tends to decrease. They confirmed this for citation and affiliation graphs extracted for patents registered with the United States Patent and Trademark Office. Contrary to this, Bettencourt, Kaiser, and Kaur (2009) showed that collaboration graphs in several scientific and technological fields exhibit initial rapid growth in their diameter, which then tends to stabilize and stay approximately constant at 12~14. This might be caused by the fact that when a new field emerges, authors are not yet aware of all relevant experts and works; as the field matures, important collaborations come into existence and lines of research are interlinked via co-author and citation linkages. The diameter of a collaboration network has major implications for information diffusion—the shorter a pathway of co-author linkages that connects an author pair, the more likely knowledge diffuses.

In our work, the collaboration network diameters seemed to stabilize at about 12. Based on the theory that a collaboration graph that the density with constant or decreasing diameters suggests that a global topological transition may occur in the graph as a whole as it grows, it could be taken to mean that we could conclude that the global topological transition had happen for the global collaboration network of IB.

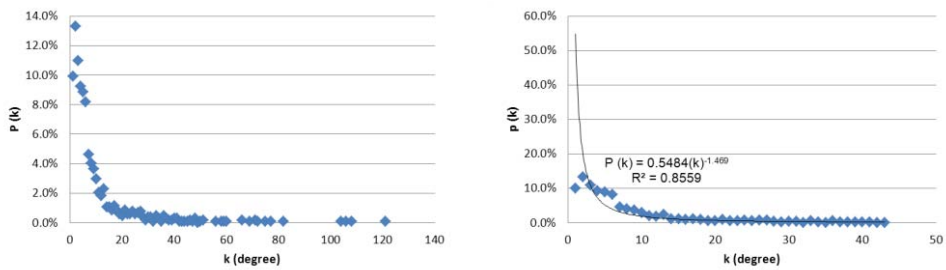
Table 1 The evolution of the patents assignees collaboration networks from core data

	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010
Network size (# Nodes)	501	670	795	874	933	977	1003	1023	1033	1037	1038
# Edges	779	1364	1919	2440	2844	3247	3700	4099	4428	4701	4893
Average degree	3.12	4.07	4.83	5.58	6.10	6.65	7.38	8.01	8.57	9.07	9.43
Density (*100)	0.62	0.61	0.61	0.64	0.65	0.68	0.74	0.78	0.83	0.88	0.91
# Components	36	26	21	21	16	16	12	14	10	10	9
Diameter	17	13	13	13	11	12	11	11	12	12	12

Distribution Patterns of Final Global Collaboration Network

During the period of 11 years development, more and more assignees had established relations with others. We mapped the relation between core nodes and their degrees in the network of 2000-2010, which was betrayed in Figure 4 (a). We could found that the number of assignees decreased with their degree increased. When nonlinear regression was used to relation between k and P(k), the

nodes in the collaboration network had degree *k* (in other words, exactly *k* links) followed a power law distribution, except for the nodes degree higher than 50. This probably not only means that the global collaboration network had the scale-free property, but also the preferential linking feature was exceeding the regression curve, there were many assignees had been chosen to link much more.



a b

Figure 4. Nodes distribution with k links (2000-2010)

Collaboration potential model strategy: Latent Collaboration Index (LCI)

The purpose of this section is to discuss the factors which might determine the possibility of patents collaborations and construct a model, Latent Collaboration Index (LCI), to evaluate this possibility. The nodes (assignees) in this paper were classified into three types of corporations, scientific establishments and universities. Although a link between two assignees could dissolution or reformation year by year, in this study we focused only on the newly established collaborations as a new link. That is to say if two assignees had collaborated in a year once before, then the link between them exist forever, no matter there was no collaboration in later years or collaborated off and on. Therefore, the new link in this paper only meant the new collaboration between assignees that have not collaborated before.

According to the results of above analysis on the evolving patterns of assignees collaboration network, the power law distribution of the network had provided the foundation for link prediction from the level of network. This section would construct a model for predicting link generation of patent assignees collaboration network based on network features together with some objective factors that might affect the link generation such as theme similarities, types of assignees, geographical distance or subjects field, etc. However, the model construction process was so a time-consuming and hard work, the demonstration of link prediction in IB was on processing now and will be discussed in later works. This paper have only presented the LCI model and discuss the factors might determine the links generation but did not carry out the demonstration process of link prediction.

LCI model

Theoretically, collaborations could be established usually by the following five ways based on the research interests or technology transfer:

Type I, randomly link occurs between two assignees both of which had no patents before. This kind of link is impossible to predict.

Type II, new assignee collaborates with the assignees that had already existed.

Type III, two assignees that both have already existed but in different network clusters.

Type IV, two assignees that both have already existed and in the same network cluster.

In the types II, III and IV, collaborations occur on a particular topology of the existing network. At the same time, because we carried out our empirical analysis on industrial biotechnology based on the patents from 2000 to 2010, there were many links occurred before 2000 and these links were marked as Type V, existing links.

Whether the network patterns would affect the choice of one assignee collaborate with others? For instance, whether the idea of preferential linking (Albert & Barabasi (2002) and Dorogovtsev & Mendes (2002)) is also appropriate for patent assignees collaboration networks? In this study we raised a hypothesis that the answers for the above queries were "yes". Furthermore, the network-based Latent Collaboration Index could be written as follows. We assumed that assignee *a* and *b* has not collaborated with each other at t-1 year, thus the possibility of *a* has been chosen to collaborate with *b* was named $LCI(a,b)_t$:

$$LCI(a,b)_t = \alpha + \beta P[d(a,b)_{t-1}] + \gamma P[k(a)_{t-1}] + \delta P[sim_{topic}(a,b)_{t-1}] + \theta\omega(a,b) + \varepsilon P[g(a,b)_{t-1}] + \Phi \quad (3)$$

Where $P[d(a,b)_{t-1}]$ represents the network distance parameter, calculated as the shortest path between two nodes in a network.

$P[k(a)_{t-1}]$ reflects the network preferential linking feature of scale-free network, is the value of degree of assignee *a*.

$P[sim_{topic}(a,b)_{t-1}]$ reflects the correlation between the link generation and the topic similarity of a pair of assignees.

$\omega(a,b)$ represents the different collaboration possibility among different types of assignees.

$P[g(a,b)_{t-1}]$ represents the correlation between the link generation and the geographical distance of a pair of assignees.

Φ represents the different collaboration tendency of different subject fields. To a given subject, Φ is a constant. This parameter is only useful for multi-subjects analysis. If an analysis is carried for a particular subject, Φ could be removed from the model, such as we did not discuss it in this paper for only IB, a particular subject had been analysis.α is a random value, which determined by the regression process.

Relativities of Factors in LCI model and link generation

(1) Degree: $P[L(a)_t] \sim P[k(a)_{t-1}]$

During the period of 2000-2010 there were 431 assignees had patents published each year and 607 new assignees. The 431 assignees had been listed descending order and divided into 9 groups with 50 assignees per group, the ninth group had only 31 assignees. Then we mapped the relation between Degree (mean) and the Compound Annual Growth Rate (mean) of each group in Fig.5. The data in the figure showed that the assignees with higher degree had also got the higher compound annual growth rate. The Pearson correlation was 0.95. It suggested that new links would like to choose assignees with higher degrees. It followed the "riches getting richer" phenomenon in the field of complex network. Applying linear regression led to the below equations:

$$P[L(a)_t] \sim 0.68P[k(a)_{t-1}] \quad (4)$$

$P[L(a)_t]$ represents the possibility of assignee *a* be chosen to collaborate in year t. $P[k(a)_{t-1}]$ represents the degree of assignee *a* in year t-1.

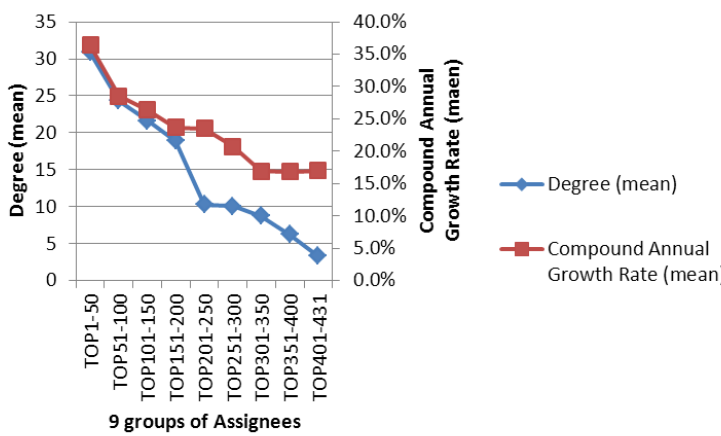


Figure 5. Degree and compound annual grow rate of assignees

Then therefore, who does became the network centre nodes? We could show them in two ways. Firstly, the assignees had been chosen to link by the 607 new assignees (table 2). Secondly, the assignees that had the highest compound annual growth rate who were the targets to be collaborated by the new assignees or the existing assignees choose to link.

The data shows that the assignees had been chosen more when the new assignees enter the network were also ranked top by degrees. There were 6 assignees in the TOP10 assignees ranked also in TOP 10 according to degree.

Another way, there were 182 linking pairs of nodes who were already existed in the network before they link to each other. The data in table 3 showed that the

assignees that had been chosen more when new link generated were also ranked top by degrees. There were 8 assignees in the TOP10 assignees ranked also in TOP 10 according to degree.

Table 2 TOP 10 assignees had been chosen to link by the 607 new assignees

	Assignees	Times be chosen	Degree (by the end of 2010)	Rank of degree
TOP1	Japan Sci & tech Agency	34	108	2
TOP2	Cnrs Cent Nat Rech Sci	15	77	6
TOP3	Dokuritsu Gyosei Hojin Sangyo Gijutsu So	14	106	3
TOP4	Inst Nat Sante & Rech Medicale	14	71	9
TOP5	Dokuritsu Gyosei Hojin Nogyo Seibutsu Sh	12	82	5
TOP6	Kyowa Hakko Kogyo Kk	9	56	17
TOP7	Inra Inst Nat Rech Agronomique	8	30	63
TOP8	Japan Min Agaric Forestry & Fisheries	8	58	16
TOP9	Univ. California	8	121	1
TOP10	Base Corp	7	36	44

Table 3 TOP 10 assignees had been chosen more when 182 links generation

		Times be chosen	Degree (by the end of 2010)	Rank of degree
TOP1	Dokuritsu Gyosei Hojin Sangyo Gijutsu So	13	106	3
TOP2	Japan Sci & Tech Agency	10	108	2
TOP3	Cnrs Cent Nat Rech Sci	7	77	6
TOP4	Dokuritsu Gyosei Hojin Rikagaku Kenkyush	7	75	7
TOP5	Us Dept Health & Human Services	7	104	4
TOP6	Inst Nat Sante & Rech Medicale	6	71	10
TOP7	Univ. California	6	121	1
TOP8	Dokuritsu Gyosei Hojin Nogyo Seibutsu Sh	5	82	5
TOP9	Du Pont De Nemours & Co E I	5	34	50
TOP10	Daiichi Pharm Co Ltd	4	36	44

(2) Distance: $P[L(a)_t] \sim [d(a, b)t-1]$

Among these 182 pair assignees, although they had already existed in the network before their links generation, there were two ways for them to construct links. One was a new link generated in one cluster and there were 169 pairs. Another way was between different clusters and two different clusters merged into one cluster with the new link generation, there were 13 pairs of new link generated from this way. These 13 pairs of assignees usually had higher values of Betweenness and were very key steps of the network evolving. While those 169 pairs contributed a large part of the new pair generation, the shortest paths of the 169 pairs before they linked were concluded in table 4. The data illustrates that the probability of link generation grew with the decrease of the shortest distance between two assignees. The link possibility of assignee *a* and *b* $P[L(a, b)_t]$ could be reflected by the distance between them as follows:

$$P[L(a, b)t] \sim e^{-0.929[d(a,\ b)_{t-1}]} \quad (5)$$

Table 4 Distribution of shortest distance of 169 pair before link

Distance before link d （→1）	Links number
2	115
3	23
4	22
5	7
6	2

(3) Types of assignees: $P[L(a)_t] \sim \omega(a,b)$

$\omega(a,b)$ represents the different collaboration possibility among different types of assignees.

This paper divides assignees into three types: corporations, scientific establishments and universities. Table 5 illustrates the types of all the collaboration pairs. We could find that "corporation- corporation" pairs had the highest capabilities of 46.4% and "university-university" pairs collaborated at least. It meant that, for instance, if there was a link generated in the collaboration network, which had a chance of 46.6% between two corporations and only 5.4% chance between two universities.

Table 5 Types of pairs of assignees

	Number of pairs	[ω(a,b)]
corporation- corporation	294	46.4%
corporation- scientific establishment	87	13.7%
corporation- university	57	9.0%
scientific establishment- scientific establishment	96	15.1%
scientific establishment- university	66	10.4%
University- university	34	5.4%

(4) Geographical distance: $P[L(a)_t] \sim P[g(a,b)_{t-1}]$

Table 6 lists the TOP 10 assignees by degrees and their collaborators' countries, which showed that the probability of they collaborated with overseas countries was only 2.1%. A great part of links were among the same countries. Thus, when an assignee had a chance to choose a partner to collaborate, the possibility of which choose an assignee from the same country was about 98%.

Therefore, the link possibility of assignee a and b $P[L(a, b)_t]$ could be reflected by the geographical distance between them as follows:

$$P[L(a, b)_t] \sim P[g(a,b)_{t-1}] \begin{cases} 97.9\% \text{ (same countries)} \\ 2.1\% \text{ （different countries）} \end{cases} \quad (6)$$

Table 6 Countries distribution of the collaborators of TOP10 assignees

Assignees	# patents	Degree	Countries	Countries distribution of the collaborators			
				USA	Japan	France	Germany
Univ. California	237	121	USA	13			
Japan Sci & Tech Agency	344	108	Japan		20		
Dokuritsu Gyosei Hojin Sangyo Gijutsu So	362	106	Japan		25		
Us Dept. Health & Human Services	332	104	USA	16		1	1
Dokuritsu Gyosei Hojin Nogyo Seibutsu Sh	300	82	Japan		14		
Cnrs Cent Nat Rech Sci	378	77	France	1		26	
Dokuritsu Gyosei Hojin Rikagaku Kenkyush	304	75	Japan		23		
Univ. Osaka	130	72	Japan		5		
Inst Nat Sante & Rech Medicale	219	71	France			14	
Univ Kyoto	139	71	Japan		4		

(5) Topic similarity: $P[L(a)_t] \sim P[sim_{topic}(a,b)_{t-1}]$

Through topic analysis on the pairs of link we found that both assignees of the link pair usually had higher topic similarity, however it could not be reverse mapping, that is to say a pair of assignees had higher topic similarity were not necessary to get link to each other. Actually, through our analysis the share of un-link pairs with higher topic similarities was very large. Not only did we discuss this problem in subject IB, a relative narrow field in the scientific world, we had also compared the topic similarities among the sub-institutes of Chinese Academy of Sciences (CAS), which had more wide subjects. The same results were found. Therefore, it might be very hard to predict the link potential although a link usually had higher topic similarity.

The methods we had used for count the topic similarities included Pearson Correlation, a common parameter used for count correlation, and two algorithms developed by ourselves. One of which is W1, which means the number of IPC that both the pair assignees shared. Another is W2, which represents the similarity of the ratios of common IPCs compared to the total IPCs for each assignee.

$$W1=\text{count}\ (IPC_i \cap IPC_j) \tag{7}$$

$$W2=\frac{\frac{N_i(IPC_i \cap IPC_j)}{N_i} \bullet \frac{N_j(IPC_i \cap IPC_j)}{N_j}}{\frac{N_i(IPC_i \cap IPC_j)}{N_i} + \frac{N_j(IPC_i \cap IPC_j)}{N_j}} \tag{8}$$

Let N be the set of independent assignees in the R&D network for a given year, for two companies i and j ($i \in I$, $j \in I$), the set of IPCs owned by i and j is represented as IPC_i and IPC_j, thus, $IPC_i \cap IPC_j$ represents the set of IPCs have been shared by i and j.

$N_i(IPC_i \cap IPC_j)$ and $N_j(IPC_i \cap IPC_j)$ represent the patent numbers of i and j belong to $IPC_i \cap IPC_j$, N_i and N_j represent the total patent numbers of i and j.

Besides these methods, there are more complicated topic models could be used for analysing the topics of assignees, then the results could be applied for count $P[sim_{topic}(a,b)_{t-1}]$. Such representative topic models included LDA (Blei, Ng & Jordan, 2003) and SAO model (Yoon & Kim, 2011). However, it is uncertain now whether these topic algorithms could be more powerful to distinguish the topics difference which might affect the link generation. More works need to be carried to make more comprehensive analysis.

To sum up, it must be explained that these factors contributing to the link generation sometimes intertwine to form an organic whole and thus become more powerful than any one of them for link prediction. Without doubt that were some other factors had not been mentioned. Even for these discussed factors, sometimes are not appropriate and should be removed from the LCI model, such as topic factor would not be used in the model until a better algorithm had been found which was powerful to identify the topic difference relative to link generation. Therefore, in this case of IB, the LCI model could be modified as follows.

$$LCI(a,b)_t = \alpha + \beta P[d(a,b)_{t-1}] + \gamma P[k(a)_{t-1}] + \theta\omega(a,b) + \varepsilon P[g(a,b)_{t-1}] \quad (9)$$

Conclusions & Discussion

The goal of the present paper is to show readers that the evolving patterns of patent assignees' collaboration evolution networks, which is the theoretical basis of constructing a model of LCI for predicting the collaboration probability between two assignees. The demonstration process was carried on the patents of field of industrial biotechnology (IB) from 2000 to 2010. The results show that during the period of 11 years development, more and more assignees had begun to apply patents (cumulative number of assignees curve in figure 2) with an growth rate of 17.3%, however the patent numbers of IB had decreased slowly year by year after it reached peak in 2002 and 2003 (figure 1), the same to the annual data of assignees in figure 2. It suggested that there were about one fifth assignees each year applied patents occasionally or discontinuously, which might be isolates or collaborated with other assignees. After 11 years accumulation, these discontinuous assignees grew to be a huge set, but only acted as the subordinate assignees in the whole set of IB assignees. At the network level, these subordinate nodes usually had very few patents or single links with other nodes or as isolates in the network. Their contributions to the network evolving were feeble. Therefore, this paper selected the assignees which had more than 50 patents and their collaborators with more than 5 patents as core data to analysis the network evolving patterns.

The scaling exponent (α=2.46) analysis had shown that the collaborations became more frequent with a higher growth rate than the assignees. It reflected more links had generated in the network as a result of the more and more assignees had sought collaborations for their research and development activities or transfer of their patented technologies in the field of IB. This could also be reflected by the evolving patterns of network size, average degree, density, number of components and diameter, etc. All these parameters had shown that the IB assignees

collaboration network had become denser and denser. Especially from the diameter we might also conclude that the IB field had come into a mature mode after finishing the topological transition occurred in about 2002 or 2003. It probably means that the group theory (Whitfield, 2008) does not only suitable for authors but could also be applicable for patents assignees, which also show the feature of increasing collaborations for the need to reinforce their innovation capabilities or to promote the technology transfer of their patents.

The final collaboration network had a scale-free pattern that included nine components. It implied a preferential linking feature of the network evolving and thus provided a foundation for link generation from the aspect of network evolving. Basing on this, two network-related parameters had been brought into the LCI model, which were degree and distance. This paper's analysis has shown the relativity of degree and link generation was positive, while the distance's relativity was negative. That is to say a link is more likely to generate between two assignees that with shorter distance and the assignees had more links would also like to be chosen. Beside the degree and distance from the point of network evolving, there are some other parameters might contribute for the link generation, such as types of assignees, geographical distance, topic similarity, etc. Different type of assignees had different probability to be linked, such as corporation had been collaborated more frequently, universities ranked lowest. When two assignees from different countries were candidate for another target assignee to link, the candidate from the same country had much more possibility to be chosen. It has to be explained that although this paper constructed the LCI model comprised topic similarity factor, it had difficulty to use it for link prediction of assignees collaboration networks at present because of the poor inference from topic similarity to collaboration.

Actually, link prediction work is impossible to get perfect, the only can we do is to increase the accuracy continually. During the improvement process, the factors in the model are needed to be adjusted or the weights of each factor need also be revised frequently according to different fields, time or data. For instance, the geographical distance factor could be refined at states or cities level, which might be more powerful. The next step of our study would use the patent data of IB from 2000 to 2005 as training data to carry the demonstration of LCI to estimate the weights of each factor. Then the validation process would predict the link generation in the field of IB from 2006 to 2010. After that, we would summarize the limitations of this LCI model and design new improvement measures to modify the LCI model step by step.

The present study leads to a necessary thinking that what had made the above characters? In SNA, although the relationships between nodes become the first priority and individual properties were only secondary. However, it should be pointed out that individual characteristics as well as relational links are necessary in order to fully understand social phenomena (Otte, 2002). In future, the node centrality and structural holes should also be analysed. Sternitzke, Bartkowski & Schramm (2008) had found that inventors spanning bridges between different

inventors groups hold patents that were technologically broader, i.e. possess more IPC classes. Whether this character could be found in assignees networks? Whether the assignees locate in central or had structural holes are more innovative or interdisciplinary?

On the other hand, our work had analysed the evolution network mainly by cumulative data. It was powerful for us to disclose the general pattern of patent assignee collaboration networks. However, some collaborators occurred only one or limited times, there were very few collaborators were related to each other all the time. Thus, if we want to look into the detail behaviours of collaborations, we need to map the collaboration networks year by year based on the annual data, not the cumulative data.

References

Albert, R. & Barabasi, A. L. (2002). Statistical mechanics of complex networks. *Reviews of Modern Physics*, 74(1): 47-97.

Bettencourt, L. M. A., Kaiser, D. I., & Kaur, J. (2009). "Scientific discovery and topological transitions in collaboration networks." *Journal of Informetrics*. 3(3): 210-221.

Blei, D. M., Ng, A. Y. & Jordan, M. I. (2003). Latent Dirichlet allocation. *Journal of Machine Learning Research*. 3(4-5): 993-1022.

Dou, H. & Bai, Y. (2007). A rapid analysis of Avian Influenza patents in the Esp@cenet database-R&D strategies and country comparisons. *World Patent Information*, 29, 26-32.

Dorogovtsev, S. N.& Mendes, J. F. F. (2002). Evolution of networks. *Advances in Physics*, 51(4): 1079-1187.

Ernst H. (2003). Patent information for strategic technology management. *World Patent Information*, 25(3): 233-242.

Gress B. (2010). Properties of the USPTO patent citation network 1963-2002. *World Patent Information*, 32(1): 3-21.

Guns R. (2011). Bipartite networks for link prediction: Can they improve prediction performance? Proceedings of ISSI 2011. 4-7, July, 2011, Durban, South Africa. - 13TH INTERNATIONAL CONFERENCE OF THE INTERNATIONAL SOCIETY FOR SCIENTOMETRICS AND INFORMETRICS, VOL1, 2011:249-260.

Hanaki, N., Nakajima, R. & Ogura, Y. (2010). The dynamics of R&D network in the IT industry. *Research Policy* 39(3): 386-399.

Han, Y.-J. & Park, Y. (2006). Patent network analysis of inter-industrial knowledge flows: The case of Korea between traditional and emerging industries. *World Patent Information*. 28(3): 235-247.

Hung, S. W. & Wang, A. P. (2010). Examining the small world phenomenon in the patent citation network: a case study of the radio frequency identification (RFID) network. *Scientometrics*. 82(1): 121-134.

Lee, S. & Kim, M.-S. (2010). Inter-technology networks to support innovation strategy: An analysis of Korea's new growth engines. *Innovation:*

Management, Policy & Practice, **12**(1: Network Analysis Application in Innovation Studies): 88-104.
Leskovec, J., Kleinberg, J. & Faloutsos, C. (2005). "Graphs over Time: Densification Laws, Shrinking Diameters and Possible Explanations" ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD). 177-187.
Linton K., Stone P. & Wise J. (2008). Patenting trends & innovation in industrial biotechnology. *Industrial Biotechnology*, *4*(4):367-390.
NWB Team. (2006). "Network Workbench Tool," Indiana University, Northeastern University, and University of Michigan. http://nwb.slis.indiana.edu/.
Otte, E. & Rousseau, R. (2002). Social network analysis: a powerful strategy, also for the information sciences. *Journal of Information Science*, 28(6): 441-453.
Segev, A. & Kantola, J. (2012). Identification of trends from patents using self-organizing maps. *Expert Systems with Applications*. 39(18): 13235-13242.
Sternitzke, C., Bartkowski, A. & Schramm, R. (2008). Visualizing patent statistics by means of social network analysis tools. *World Patent Information*, 30(2): 115-131.
Wartburg, I., Teichert, T. & Rost, K. (2005). Inventive progress measured by multi-stage patent citation analysis. *Research Policy*, 2005, 34(10): 1591-1607.
Whitfield, J. (2008). Group theory. *Nature*, 455, 720-723.
Yang, Y. Y., Akers, L., Yang, C. B., *et al*. (2010). Enhancing patent landscape analysis with visualization output. *World Patent Information*. 32(3): 203-220.
Yoon, J. & Kim, K. (2011). Identifying rapidly evolving technological trends for R&D planning using SAO-based semantic patent networks. *Scientometrics*. 88(1): 213-228.

MATCHING BIBLIOGRAPHIC DATA FROM PUBLICATION LISTS WITH LARGE DATABASES USING N-GRAMS (RIP)

Mehmet Ali Abdulhayoglu[1] and Bart Thijs[2]

[1] *mehmetali.abdulhayoglu@kuleuven.be*
Centre for R&D Monitoring (ECOOM), K.U. Leuven, Waaistraat 6, B-3000 Leuven (Belgium)

[2] *bart.thijs@kuleuven.be*
Centre for R&D Monitoring (ECOOM), K.U. Leuven, Waaistraat 6, B-3000 Leuven (Belgium)

Abstract
This paper presents a matching method for the identification of publications, extracted from publications lists provided by authors or research institutes, in large bibliographic databases. For this purpose, Levenshtein similarities based on N-grams have been used to measure the closeness between the given publications and the database records. Several different similarity scores have been calculated and used as variables in a kernel discriminant analysis. About 95% accuracy has been achieved by using this method.

Conference Topic
Management and Measurement of Bibliometric Data within Scientific Organizations (Topic 9).

Introduction
The application of evaluative bibliometrics at the micro and meso level often requires the use of CVs and publication lists provided by authors, applicants and institutions as an input of the analysis. The quality of the retrieved data sets is often a crucial determinant of the validity of the final results. Therefore, the identification of the authors' or research teams' publications in large databases like the Web of Science (WoS) and Scopus usually results in a tremendous amount of manual effort. Automation of this process of directly linking publications provided in lists to publication records indexed in the database could essentially simplify this task and free up resources previously assigned to the manual cleaning tasks.
Most commonly, the problem in finding such matches is caused by incomplete, erroneous or censored data in publication lists; but erroneous entries occur in the databases as well. It is very likely that an author adds a publication to his/her CV as soon as a paper is submitted or conditionally accepted but later, the actual publication year, but also the title or the number and sequence of co-authors can change during the process of revision and finalization of the publication.

In the present paper, a promising method is based on measuring complete string similarity between the bibliographic references in both publication lists and the bibliographic database is presented. In the literature, there are numerous works related to reference/citation matching for many different purposes (e.g. Giles et al. (1998), Lawrence et al. (1999), Larsen (2004)). Depending on implementation, a proper similarity measure has been questioning for string comparison. Kondrak (2005) introduced a modified Levenshtein distance which is based on N-grams that is applied in this study.

Possible matches with publications indexed in the database have been examined by kernel discriminant analysis (KDA) which uses various similarity scores as variables. In particular, we analyze different components, such as title, journal name, co-authors, etc. that constitute the variables in KDA and the accuracy of matches.

The objective of this study is not to completely automatize the match or to definitely decide whether the given publication is indexed in the database but to enhance the matching accuracy by trying to obtain various variables to find an optimum solution for discriminating existing and non-existing entries. By achieving this objective, it is intended to reduce manual work to the possible minimum.

In order to test the efficiency of the proposed method, a publication list taken from scientists' CVs has been matched with papers indexed in the WoS database. The results show that the model proposed by linear discriminant function is capable of matching the publications in our test set with almost 95% accuracy.

First we introduce the concept of N-grams and how that is implemented in this application. We will also describe the classification model that is built. In the subsequent section, the data is described and results are discussed.

Methodology

N-gram

A character N-gram is an adjacent sequence of *n* characters from a given text. The key benefit of using N-gram is that textual errors can be handled since each string is decomposed in small chunks of text. If an error occurs, this is included in only a minor number of chunks and all others remains intact. Moreover, one needs not to deal with stemming since the corresponding forms of a word (e.g. 'search', 'searching', 'searched') have a lot in common as it is decomposed into their N-grams (Cavnar and Trenkle, 1994).

Based on the idea in Kondrak (2005), we have used the 'NGramDistance' implementation which is part of the LUCENE software for string comparison. This implementation is in fact an N-gram version of the Levenshtein-Edit distance. This distance describes the minimum number of single-character edits that have to be made in order to change one string to another (Levenshtein, 1966). For the N-gram based edit distance between strings *x* and *y*, a matrix $M_{1\ldots m+1,1\ldots n+1}$ is constructed where $M_{i,j}$ is the minimum number of edit

operations needed to match $x_{1\ldots i}$ to $y_{1\ldots j}$. Each matrix element $M_{i,j}$ is calculated according to Eqs. (1) – (3), where 'cost' is the total number of distinct letters according to their positions between the N-grams x_i, y_j and n is the size of the N-gram.

$$M_{1,1} = 0 \tag{1}$$

$$M_{i,j} = min\begin{cases} M_{i-1,j} + 1 \\ M_{i,j-1} + 1 \\ M_{i-1,j-1} + \delta(x_i, y_j) \end{cases} \tag{2}$$

$$\delta(x_i, y_j) = cost/n \tag{3}$$

To see how the N-grams are compared to calculate $\delta(x_i, y_j)$, we can examine the following example. Assume that 3-grams partitions are being used and the underscores indicate the blanks. The 'cost' for "_th" and "_th" 3-grams comparison is 0 while it is 3 for "_th" and "dif" since there is no common letter at the same position at all. Nevertheless, the 'cost' between the portions "ff." and "ffu" is 2 since there are two same letters at the same positions and $\delta(x_i, y_j)$ is 2/3.

Here it should be mentioned that we use 3-grams considering the length of the components on CVs. Since such components as author names or publication year contain short texts, there is no need to use N-grams in big sizes to grab a straightforward similarity.

The suggested Levenshtein distance has two features. It normalizes the original distance measure by the length of the longer string to avoid length bias. In addition, it adds a null-character prefix of size *n*-1 such that the initial letter is contained in the same number of N-grams and can be exploited more efficiently since the initial characters are very important in word similarity. It should be stressed that null-character prefix matches are discounted so that strings with no matching characters will return maximum distance. Finally, distances are subtracted from 1 to get the similarity scores ranging between 0 and 1, where 1 or 0 means that the specified strings are identical or maximally different, respectively.As we can expect that publication lists provide detailed bibliographic information about the publications such as its title, the journal title, the names of the author and co-author(s), publication year, volume and first and end page. However, it is also very likely that records in the lists do not contain complete information or do not follow the standards of the database. Therefore, it is necessary to compare each entry in the publication lists with a set of variations from the database records. For each entry in the publication list and the different variations of the database records, the similarity scores are calculated. Table 1 gives an overview of all eight variations that were used with the order of components. The numbers given in the table indicate the order of the components

to be used to construct the references from the database records. For each variation the corresponding similarity score for the publication in publication list is calculated. Finally the maximum of these similarity scores is also assigned to SCORE9 variable for our analysis.

Table 1: Components in Database and corresponding scores

Components / Variables	Title	Journal	Co-authors	Volume	Begin Page	Publication Year
SCORE1	2	3	1	4	5	6
SCORE2	2	-	1	3	4	5
SCORE3	1	2	-	3	4	5
SCORE4	1	-	2	-	-	-
SCORE5	1	-	-	-	-	-
SCORE6	1	2	-	-	-	-
SCORE7	-	1	2	-	-	-
SCORE8	-	1	-	-	-	-

Table 2 shows an example of 3-gram similarity scores. One entry from a publication list is matched with six different versions of the same paper by using its components indexed in the Web Of Science database.

Table 2: CV and Variable Similarity Scores with 3-grams [Data sourced from Thomson Reuters Web of Knowledge]

CV	Zhang, L., Thijs, B., Glänzel, W., The diffusion of H-related literature. JOI, 2011, 5 (4), 583-593	**Variables**	**3-Gram Score**
WOS	ZHANG, L, THIJS, B, GLANZEL, W, The diffusion of H-related literature, JOURNAL OF INFORMETRICS, 5, 583, 2011	SCORE1	0,67
WOS	ZHANG, L, THIJS, B, GLANZEL, W, The diffusion of H-related literature, 5, 583, 2011	SCORE2	0.73
WOS	The diffusion of H-related literature, JOURNAL OF INFORMETRICS, 5, 583, 2011	SCORE3	0.34
WOS	ZHANG, L, THIJS, B, GLANZEL, W, The diffusion of H-related literature	SCORE4	0.65
WOS	The diffusion of H-related literature	SCORE5	0.36
WOS	The diffusion of H-related literature, JOURNAL OF INFORMETRICS	SCORE6	0.41
		SCORE9	**0,73**

Discriminant Analysis (DA)

A classification model is made by using Discriminant Analysis (DA). This model enables us to automatically classify a match between a publication in the CV and an entry from the database as being correct on the basis of the set of N-Gram

scores. DA requires two important assumptions, namely, multivariate normality and homogeneity of variances/co-variances. However, it is observed that none of the assumptions are hold. Therefore, we applied the non-parametric discriminant analysis, namely, Kernel Discriminant Analysis (KDA) which is a powerful learning technique (Taylor and Cristianini, 2004). KDA is a method which implements a linear discriminant analysis in a feature space. That is, a non-linear mapping in the input space is handled by a linear mapping by means of a kernel function. It is based on estimating a nonparametric density function for each group in the training set and present a classification criterion (SAS 9.2 User Guide, 2008). As deriving the best classification results, we have decided to use *normal* kernel (with mean 0 and variance r^2V_t), after trying several other kernels such as Uniform, Epanechnikov, Biweight and Triweight. Here, V_t is the variance-covariance matrix for group t and r is the smoothing parameter to determine the degree of irregularity in the estimate of density function. Within-group covariance matrixes are used instead of pooled covariance matrixes in the analysis since the variability between groups are significantly different. Here, deciding the r is the most crucial part of the analysis. As stated in Khattree and Naik (2000), trying various values of r and choosing the one working best is the solution.

We have implemented KDA for the two sets of variables. The first set comprises the variables SCORE1, SCORE2, SCORE3, SCORE4, SCORE5 and SCORE6 while the second set comprises SCORE9, SCORE5 and SCORE8. While the former set is chosen to examine the variables all including "Title" component and its variations, the latter one is chosen to analyse as a relatively more independent set with "Maximum", "Title" and "Journal Name".

This resulted in different training models. Based on these models with the estimated group-specific densities, the classifiers could be retrieved that assign publication pairs in the test set to one of the two groups.

Data

Two large samples of entries from a real world application have been collected. The samples were taken from CV's and matched with publication lists. These pairs of reference and publication were separated into training and test sets in order to be able to validate the classification model. The first set, training set consists of 6525 publications-reference pairs and all these references were matched manually to an entry in the WOS-database. Matching this data manually took several days in manpower. These pairs are labelled as members of Group 1. A second group (Group 0) was created by randomly choosing three unmatched records from the database for each of the 6525 references. Thus 19575 pairs belong to the Group 0. In addition, we have 2570 pairs in a test set to be classified into Group 1 or Group 0 by using the results from the training set. The publications in the test set are completely different from those in the training set.

Results

Using the *normal* kernel with the smoothing parameter r=3 for the set2 variables, much better results are observed than the *normal* kernels with other r values and than the other kernels. Table 3 presents the percentages of the classification accuracies and the errors for the different values of the smoothing parameter according to the set1 and set2 variables when using *normal* kernel. We define the publications assigned to Group 0 incorrectly as *false negatives* and the ones assigned to Group 1 incorrectly as *false positives*. It is obvious that the percentage of *false positives (false negatives)* is decreasing (increasing) as r increasing. We can infer that there is a trade-off between these errors and the most balanced case is observed for the set2 with r=3 as having the highest accuracy. Here, it should be noted that the smaller number of *false positives* with a high accuracy is more important for us since we deal with a sample database. That is, a new publication might not be matched with the publications indexed in the database although it exists in WoS. Therefore, our proposed model is the *normal* kernel with r=3 for the second variable set.

Table 3: Classification accuracy and error percentages for the different values of smoothing parameter according to the variables in Set1 vs. Set2 [Data sourced from Thomson Reuters Web of Knowledge]

set1	*r*	2	3	4	5
	accuracy	92.96	90.3	78.25	60.62
	false negatives	4.48	13.9	33.31	60.53
	false positives	6.20	1.20	0.20	0
set2	*r*	2	3	4	5
	accuracy	91.13	94.9	90.97	82.33
	false negatives	2.63	5.68	13.33	27.03
	false positives	10.15	2.23	0.62	0.16

Table 4 summarizes the classification results according to the proposed model for the publications in the test set. The rows present their observed groups while the columns present their estimated groups. According to the results, 94.3% of the publications belong to Group 1 is classified correctly by the proposed model. Furthermore, 97.8% of the publications which is estimated as in Group 1 are classified correctly.

Table 4: Classification results according to the proposed model [Data sourced from Thomson Reuters Web of Knowledge]

	Estimated 0	Estimated 1	Total
Observed 0	862	36	898
Observed 1	95	1555	1672
Total	957	1613	2570

Figure 1 visualizes the classification problem. It contains two diagrams which present the observed matching group values (left panel) and estimated matching group values (right panel) for the publications from the CV lists in the test set. The diagram (right) shows that the vast majority of the publications in Group1 is classified correctly. However, with the proposed model, *false positives* remain as an issue as it is seen through the figure.

Here, it is important to note that maximum matching scores below 0.30 have been omitted for the test set. Hence, we have aimed to enhance the performance of the model while implementing it to the whole database.

At this point, it should be mentioned that, in principle, it is possible to find some *false positive* matches based on high similarity scores, which, in fact, do not cover the same publication. However, in both the training and test sets, such cases did not occur. Therefore this situation requires further investigation.

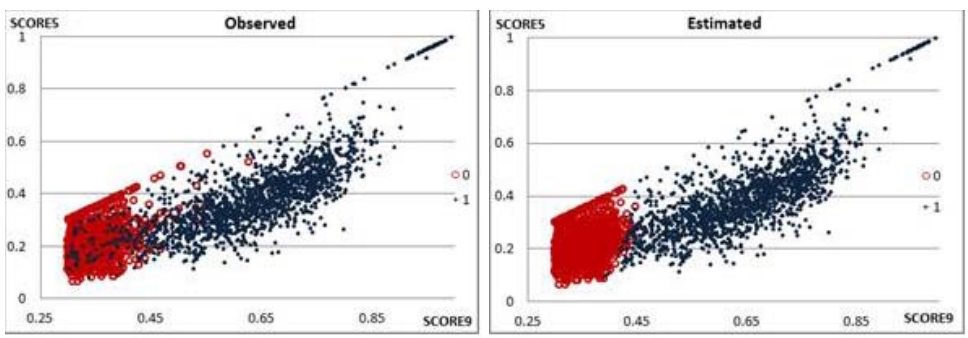


Figure 1: The Scatter Plots for Observed (Left) and Estimated (Right) Group Values in the Test Set According to the Most Powerful Discriminator Variables (SCORE9 vs. SCORE5) Exist in the Proposed Model [Data sourced from Thomson Reuters Web of Knowledge]

Conclusion

Almost 95% of the publications collected from a CV's publication list could be correctly matched with the Web of Science database using the proposed model. However, there are some issues that should be mentioned here. Firstly, a sufficiently large publication sample of the WoS has been used. Although this approach already provides important insight, the aim should be searching for the publications in complete database. Secondly, *false positives* remain as an issue. Normally, these CV publications belong to Group 0, yet the model assigns them to Group 1.The tolerance of *false positives* depends on the application. *False positive* or *false negative* drawbacks could be overcome using regular expressions. Extracting content-relevant information from publication strings, could possibly improve the accuracy of matching. Eventually, the method applied in this paper can especially be leveraged for the evaluation purposes such as job promotion cases at micro level or as macro assessment studies, that large number of CVs are to be dealt with.

References

Cavnar, W.B. & Trenkle, J.M. (1994). *N-Gram-Based Text Categorization*. In: Proceedings of SDAIR-94, 3rd Annual Symposium on Document Analysis and Information Retrieval, Las Vegas, US, 161–175.

Khattree , R. & Naik, D.N. (2000). Multivariate Data Reduction and Discrimination With SAS Software. *Cary, NC: SAS Institute Inc.* Wiley.

Kondrak, G. (2005). N-gram similarity and distance. *Proceedings of the Twelfth International Conference on String Processing and Information Retrieval (SPIRE 2005)*, pp. 115-126, Buenos Aires, Argentina.

Levenshtein, V.I. (1966). Binary codes capable of correcting deletions, insertions and reversals. *Soviet Physics-Doklady*, 10, 707-710.

Shawe-Taylor, J. & Cristianini, N. (2000). *Kernel Methods for Pattern Analysis*. Cambridge Univ. Press.

SAS Institute Inc. (2008). *SAS/STAT 9.2 User's Guide*. Cary, NC: SAS Institute Inc.

Lawrence, S., Giles, C.L. and Bollacker, K.D. (1999*). Autonomous Citation Matching*. In: Etzioni, O., Muller, J.P. and Bradshaw, J.M. eds. AGENTS '99. Proceedings of the Third Annual Conference on Autonomous Agents, May 1-5, 1999, Seattle, WA, USA. New York: ACM Press, p.392-393.

Larsen, B. (2004). *References and Citations in Automatic Indexing and Retrieval Systems - Experiments with the Boomerang Effect*. PhD thesis, Royal School of Library and Information Science.

Giles, C.L., Bollacker, K.D., Lawrence, S. (1998). *CiteSeer: an automatic citation indexing system*. Digital 98 Libraries. Third ACM Conference on Digital Libraries. Pg. 89-98.

MATHEMATICAL CHARACTERIZATIONS OF THE WU- AND HIRSCH-INDICES USING TWO TYPES OF MINIMAL INCREMENTS

Leo Egghe

leo.egghe@uhasselt.be
Universiteit Hasselt, Campus Diepenbeek, Agoralaan, B-3590 Diepenbeek (Belgium)
Universiteit Antwerpen, Stadscampus, Venusstraat 35, B-2000 Antwerpen (Belgium)

Abstract
For a general increasing function $f(n)$ $(n=1,2,3,...)$ we can define the most general version of the Hirsch-index being the highest rank n such that all papers on ranks $1,...,n$ each have at least $f(n)$ citations. The minimum configuration to have this value of n is n papers each having $f(n)$ citations, hence we have $nf(n)$ citations in total. To increase the value n by one we hence need (minimally) $(n+1)f(n+1)$ citations, an increment of $I_1(n)=(n+1)f(n+1)-nf(n)$ citations. Define the increment of second order as $I_2(n)=I_1(n+1)-I_1(n)$. We characterize the general Wu-index by requiring specific values of $I_1(n)$ and $I_2(n)$, hence also characterizing the Hirsch-index.

Conference Topic
Scientometrics Indicators (Topic 1)

Introduction
The most general Hirsch-type index can be defined by using a general increasing function $f(n)$ $(n=1,2,3,...)$. The definition is as follows. Let us have a set of papers where the i[th] paper has c_i citations (i.e. received c_i citations). We assume that papers are arranged in decreasing order of received citations (i.e. $c_i \geq c_j$ if and only if $i \leq j$). The most general Hirsch-type index can be defined as the highest rank n such that all papers on ranks $1,...,n$ have at least $f(n)$ citations. Well-known examples are $f(n)=n$ for the classical Hirsch-index (h-index), Hirsch (2005), $f(n)=an$ $(a>0)$ for the general Wu-index (Egghe (2011) and Wu (2010) for $a=10$), $f(n)=n^a$ $(a>0)$ for the general Kosmulski-index (Egghe (2011) and Kosmulski (2006) for $a=2$). Note that the general Wu- and Kosmulski-indices reduce to the h-index for $a=1$.

It is important, at least from a theoretical point of view, to know for these h-type indices, how (e.g.) an author can increase his/her h-type index value from *n* to $n+1$ (for any *n = 1,2,...*). In other words, it is important to know what effort is required from an author to increase his/her h-type index by one.

In general $c_i \geq f(n)$ for $i = 1,...,n$ but in many cases we will have $c_i > f(n)$. However the minimum situation to have an index equal to n is to have n papers with exactly $f(n)$ citations each and where the other papers have zero citations. In this case we have a total of $nf(n)$ citations. To have the minimal situation for an index equal to $n+1$, we need $n+1$ papers with exactly $f(n+1)$ citations each and where the other papers have zero citations. Now we have a total of $(n+1)f(n+1)$ citations. We define the general increment of order 1 as, for every n:

$$I_1(n) = (n+1)f(n+1) - nf(n) \tag{1}$$

The general increment of order 2 is defined as

$$I_2(n) = I_1(n+1) - I_1(n) \tag{2}$$

which is equal to, by (1)

$$I_2(n) = (n+2)f(n+2) - 2(n+1)f(n+1) + nf(n) \tag{3}$$

Examples:

1. For the general Wu-index ($f(n) = an$) we have

$$I_1(n) = a(2n+1) \tag{4}$$

$$I_2(n) = 2a \tag{5}$$

for all n, as is readily seen.
This gives for the h-index:

$$I_1(n) = 2n+1 \tag{6}$$

$$I_2(n) = 2 \tag{7}$$

for all n.

2. For the general Kosmulski-index ($f(n) = n^a$) we have

$$I_1(n) = (n+1)^{a+1} - n^{a+1} \tag{8}$$

$$I_2(n) = (n+2)^{a+1} - 2(n+1)^{a+1} + n^{a+1} \tag{9}$$

for all n.

3. For the threshold index (obtained for $f(n)=C$, a constant) (called the "highly cited publications indicator" in Waltman and van Eck (2012)) we have

$$I_1(n)=C \tag{10}$$

$$I_2(n)=0 \tag{11}$$

for all n.

In the next section we will characterize the functions $f(n)$ for which (4) is valid. It turns out that we obtain a class of functions much wider than $f(n)=an$ and from this we will characterize the general Wu-index. From this we will also obtain a characterization of the h-index. The same will be done for the threshold index.

In the third section we will characterize the functions $f(n)$ for which (5) is valid. Again it turns out that we obtain a class of functions much wider than $f(n)=an$ and from this we will newly characterize the general Wu-index. From this we will also refind a characterization of the h-index, already proved in Egghe (2012).

The paper ends with a conclusions section and with suggestions for further research.

Characterization of functions $\boldsymbol{f(n)}$ that satisfy $\boldsymbol{I_1(n)=a(2n+1)}$ for all n and characterization of the Wu- and Hirsch-indices and analogue for the threshold index.

So we put, for all n,

$$I_1(n)=(n+1)f(n+1)-nf(n)=(2n+1)a \tag{12}$$

Hence

$$f(n+1)=\frac{n}{n+1}f(n)+a\frac{2n+1}{n+1} \tag{13}$$

This shows that we can choose one free parameter: $f(1)>0$. From (13) we now have

$$f(2)=\frac{1}{2}f(1)+a\frac{3}{2} \tag{14}$$

$$f(3)=\frac{1}{3}f(1)+\frac{8}{3}a \tag{15}$$

(now also using (14))

$$f(4)=\frac{1}{4}f(1)+\frac{15}{4}a \tag{16}$$

(now also using (15)).

From this mechanism we can formulate and prove the next Theorem.

Theorem 1:

$$I_1(n)=a(2n+1)$$

for all n if and only if

$$f(n)=\frac{1}{n}f(1)+\frac{n^2-1}{n}a \tag{17}$$

for all n.

Proof:
The proof is by complete induction. It is clear that (17) is valid for $n=1$ and we proved (17) for $n=2,3,4$. Now we suppose that (17) is true for n. For $n+1$ we have by (12) (hence (13))

$$f(n+1)=\frac{n}{n+1}f(n)+a\frac{2n+1}{n+1}$$

By (17) we have

$$f(n+1)=\frac{n}{n+1}\left[\frac{1}{n}f(1)+\frac{n^2-1}{n}a\right]+a\frac{2n+1}{n+1}$$

$$f(n+1)=\frac{1}{n+1}f(1)+\frac{a}{n+1}(n^2-1+2n+1)$$

$$f(n+1)=\frac{1}{n+1}f(1)+\frac{(n+1)^2-1}{n+1}a$$

which is (17) for $n+1$. Hence (17) is valid for all n.

Reversely, if we have (17), we have to show that (12) is valid. Indeed, for all n

$$I_1(n)=(n+1)f(n+1)-nf(n)$$

$$I_1(n)=(n+1)\left[\frac{1}{n+1}f(1)+\frac{(n+1)^2-1}{n+1}a\right]-n\left[\frac{1}{n}f(1)+\frac{n^2-1}{n}a\right]$$

$$I_1(n)=(2n+1)a$$

Hence (12) is valid for all n.

Note that, for $a=1$, we have a characterization of the Hirsch-type increment $I_1(n)=2n+1$ (see (6)).

From Theorem 1 we can prove a characterization of the general Wu-index.

Theorem 2:

We have equivalent of

(i) $I_1(n)=a(2n+1)$ for all n and $f(1)=a$

(ii) $f(n)=an$ for all n (i.e. we have the Wu-index)

Proof:

(i) => (ii)

By formula (17) in Theorem 1 we have for all n

$$f(n)=\frac{a}{n}+\frac{n^2-1}{n}a$$

$$f(n)=na$$

(ii) => (i)

It was already shown in the introduction that the Wu-index satisfies (12).

Note that Theorem 2 for $a=1$ yields a characterization of the Hirsch-index.

Note that $f(n)$ in (17) increases if $a\geq\frac{f(1)}{2}$:

$$f'(n)=\frac{n^2a-f(1)+a}{n^2}\geq 0$$

if and only if

$$(n^2+1)a\geq f(1)$$

for all n. It suffices to require

$$2a\geq f(1)$$

or

$$a \geq \frac{f(1)}{2}$$

Now we will prove the analogue result for the threshold index. So let $f(n) = C > 0$ for all n (C: a constant). We showed in the introduction that $I_1(n) = C$ for all n. Let us characterize all functions $f(n)$ that satisfy this. So

$$I_1(n) = (n+1) f(n+1) - n f(n) = C \tag{19}$$

for all n. Hence

$$f(n+1) = \frac{n}{n+1} f(n) + \frac{C}{n+1} \tag{20}$$

Again we use the general parameter $f(1) > 0$. We have, by (20)

$$f(2) = \frac{1}{2} f(1) + \frac{C}{2} \tag{21}$$

$$f(3) = \frac{1}{3} f(1) + \frac{2C}{3} \tag{22}$$

(now also using (21))

$$f(4) = \frac{1}{4} f(1) + \frac{3C}{4} \tag{23}$$

(now also using (22)). Hence we can formulate and prove Theorem 3

Theorem 3:

$I_1(n) = C$ for all n if and only if

$$f(n) = \frac{1}{n} f(1) + \frac{n-1}{n} C \tag{24}$$

for all n.

Proof:

The proof is by complete induction. We have already (24) for $n = 1$ and proved (24) for $n = 2,3,4$. Now we suppose (24) is valid for n. For $n+1$ we have by (20)

$$f(n+1) = \frac{n}{n+1} f(n) + \frac{C}{n+1}$$

$$f(n+1)=\frac{n}{n+1}\left[\frac{1}{n}f(1)+\frac{n-1}{n}C\right]+\frac{C}{n+1}$$

$$f(n+1)=\frac{1}{n+1}f(1)+C$$

which is (24) for $n+1$. So (24) is proved for all n.

Reversely, if we have (24) for all n, we have

$$I_1(n)=(n+1)f(n+1)-nf(n)$$

$$I_1(n)=(n+1)\left[\frac{1}{n+1}f(1)+\frac{n}{n+1}C\right]-n\left[\frac{1}{n}f(1)+\frac{n-1}{n}C\right]$$

$$I_1(n)=C$$

for all n.
From Theorem 3 we can prove a characterization of the threshold index.

Theorem 4:
We have equivalency of

(i) $I_1(n)=C$ for all n and $f(1)=C$

(ii) $f(n)=C$ for all n (hence the threshold index).

Proof:
(i) => (ii)
This is clear from (24), using that $f(1)=C$

(ii) => (i)
This was already proved in the introduction.
Note that $f(n)$ in (24) increases if and only if $C\geq f(1)$. Indeed

$$f'(n)=\frac{C-f(1)}{n^2}\geq 0$$

if and only if $C\geq f(1)$.

Characterization of functions $f(n)$ that satisfy $I_2(n)=2a$ for all n and characterization of the Wu- and Hirsch-indices and analogue for the threshold index

So we put, for all n

$$I_2(n)=(n+2)f(n+2)-2(n+1)f(n+1)+nf(n)=2a \tag{25}$$

Hence

$$f(n+2)=\frac{2(n+1)}{n+2}f(n+1)-\frac{n}{n+2}f(n)+\frac{2a}{n+2} \tag{26}$$

for all n. Hence we can choose two free parameters: we choose $f(1)$, $f(2)$. Since we only want to work with increasing functions $f(n)$ we suppose $f(2)\geq f(1)$. By (26) we have

$$f(3)=\frac{4}{3}f(2)-\frac{1}{3}f(1)+\frac{2a}{3} \tag{27}$$

$$f(4)=\frac{6}{4}f(2)-\frac{2}{4}f(1)+\frac{6a}{4} \tag{28}$$

(now also using (27))

$$f(5)=\frac{8}{5}f(2)-\frac{3}{5}f(1)+\frac{12}{5}a \tag{29}$$

(now also using (28)).

Hence we can formulate and prove Theorem 5.

Theorem 5:

$I_2(n)=2a$ for all n if and only if

$$f(n)=\frac{1}{n}\left[2(n-1)f(2)-(n-2)f(1)+(n-1)(n-2)a\right] \tag{30}$$

for all n.

Proof:

The proof is by complete induction. We already proved (30) for $n=3,4,5$ and is easy to see for $n=1,2$. Now we suppose that (30) is valid for *n* and $n+1$. For $n+2$ we have , by (25)

$$f(n+2)=\frac{2(n+1)}{n+2}\left[\frac{2nf(2)-(n-1)f(1)+n(n-1)a}{n+1}\right]$$
$$-\frac{n}{n+2}\left[\frac{2(n-1)f(2)-(n-2)f(1)+(n-1)(n-2)a}{n}\right]+\frac{2a}{n+2}$$

$$f(n+2)=\frac{1}{n+2}\left[2(n+1)f(2)-nf(1)+n(n+1)a\right] \qquad (31)$$

after an elementary calculation. Now (31) is (30) for $n+2$.
Reversely, if (30) is valid for all n, it is an elementary calculation, using (25), that $I_2(n)=2a$ for all n.
From Theorem 5 we can prove a characterization of the general Wu-index.

Theorem 6:
We have equivalency of

(i) $I_2(n)=2a$, for all n and $f(1)=a$ and $f(2)=2a$

(ii) $f(n)=na$ for all n (hence we have the general Wu index).

Proof:
(i) => (ii)
It follows from (30) in Theorem 5 that, for $f(1)=a$, $f(2)=2a$ that $f(n)=na$ for all n.

(ii) => (i)
We proved in the introduction that the Wu-index satisfies $I_2(n)=2a$ for all n.
Note that, for $a=1$, Theorem 6 is a characterization of the Hirsch-index, which appeared already in Egghe (2012).
Note: It is easy to see that $f(n)$ in (30) is an increasing function. This can be shown using (30) by calculating $f'(n)$ or by (26) using complete induction (and, in both cases, using that $f(1)\le f(2)$).
For the sake of completeness we also mention the following characterization of $I_2(n)=0$ for all n and of the threshold index.

Theorem 7 (Egghe (2012)):

$I_2(n)=0$ for all n if and only if

$$f(n)=\frac{2(n-1)f(2)-(n-2)f(1)}{n} \tag{32}$$

for all n.

Theorem 8 (Egghe (2012)):

The following assertions are equivalent:

(i) $I_2(n)=0$ for all n, $f(1)=f(2)=C$ a positive constant.

(ii) $f(n)=C$ for all n, i.e. we have the threshold index.

Conclusions and suggestions for further research

In this paper we characterized functions for which $I_1(n)=(2n+1)a$ for all n. As a consequence we proved a characterization of the general Wu-index, hence also of the h-index.

We then characterized functions for which $I_2(n)=2a$ for all n. As a consequence we proved a new characterization of the general Wu-index, hence also of the h-index.

For the threshold index we executed the same exercise leading to characterizations of the threshold index.

We invite the reader to elaborate further studies on $I_1(n)$ and $I_2(n)$, hereby characterizing other known and new impact indices. We stress the importance of such studies, at least from a theoretical point of view. Characterizing indices which require a certain increment of citations in order to increase the index with one unit shows what effort is required from the author to reach this increase.

References

Egghe L. (2011). Characterizations of the generalized Wu- and Kosmulski-indices in Lotkaian systems. *Journal of Informetrics,* 5(3), 439-445.

Egghe L. (2012). A mathematical characterization of the Hirsch-index by means of minimal increments. Preprint.

Hirsch J.E. (2005). An index to quantify an individual's scientific research output. *Proceedings of the National Academy of Sciences of the United States of America*, 102(46), 16569-16572.

Kosmulski M. (2006). A new Hirsch-type index saves time and works equally well as the original h-index. *ISSI Newsletter*, 2(3), 4-6.

L. Waltman and N.J. van Eck (2012). The inconsistency of the h-index. *Journal of the American Society for Information Science and Technology* 63(2), 406-415.

Wu Q. (2010). The w-index: A measure to assess scientific impact by focusing on widely cited papers. *Journal of the American Society for Information Science and Technology*, 61(3), 609-614.

MEASURING INTERNATIONALISATION OF BOOK PUBLISHING IN THE SOCIAL SCIENCES AND HUMANITIES USING THE BARYCENTRE METHOD (RIP)

Frederik T. Verleysen and Tim C.E. Engels

Frederik.Verleysen@ua.ac.be
Centre for Research & Development Monitoring (ECOOM), University of Antwerp, Middelheimlaan 1, 2020 Antwerp (Belgium)

Tim.Engels@ua.ac.be
Department of Research Affairs and Centre for Research & Development Monitoring (ECOOM), University of Antwerp, Middelheimlaan 1, 2020 Antwerp (Belgium); Antwerp Maritime Academy, Noordkasteel-Oost 6, 2030 Antwerp, Belgium

Abstract
Using places-of-publication barycentres this paper measures the internationalisation of book publishing in the Social Sciences and Humanities (SSH) as practiced at Flemish universities. Over a ten-year timespan, the barycentre for monographs, edited books and book chapters has moved south-westwards slightly, away from Flanders, Belgium. A comparison of 16 SSH disciplines demonstrates how European continental, British, American and other publishers carry a different weight depending on the discipline.

Introduction
This paper examines aspects of internationalisation of scholarly book publishing by researchers affiliated with Flemish universities in the period 2002-2011. We apply the concept of barycentre to the place of publication of peer reviewed monographs, edited books and book chapters. A barycentre of book publishing is defined as the imaginary point at which a flat, weightless but stiff map of the world would balance if weights of identical value were placed on it so that each weight represented the place of publication of one monograph, edited book or book chapter (Bartlett, 1985; Jin & Rousseau, 2001). Two aspects of internationalisation are analysed in particular. First, it is determined whether the places-of-publication barycentre for book publications in the Social Sciences and Humanities (SSH) has moved during the period under study. Second, the barycentres for 16 SSH disciplines are calculated with a view of visualising differences in internationalisation. The further away a barycentre is from the barycentre of the five Flemish universities, the more frequent authors have published with a publisher that is not situated in Flanders. Often this is an Anglophone, i.e. British or American publisher. The analysis of full coverage data on peer-reviewed publications in the SSH has previously shown that SSH journal articles are increasingly published in English and in the Web of Science

(Ossenblok, Engels, & Sivertsen, 2012), two major indications of internationalisation of scholarly SSH research. Adding to this, the comparison of barycentres presented in this paper offers information on aspects of internationalisation of book publishing, thereby broadening the picture of publication patterns in the SSH (Hicks, 2004). A growing geographical distance of the places-of-publication barycentres to the place of affiliation of the scholar(s) involved (i.e. the Flemish universities), and hence a smaller role for local publishers, would in its own right be indicative of growing internationalisation of scholarly book publishing.

For Flanders, comprehensive data on SSH book publications is available through the Flemish Academic Bibliographic Database for the Social Sciences and Humanities ("Vlaams Academisch Bibliografisch Bestand voor de Sociale en Humane Wetenschappen" or VABB-SHW, www.ecoom.be/en/vabb). This database was established in 2008-2010 in order to include the peer reviewed SSH publications in the regional performance-based research funding system (PRFS) (Engels, Ossenblok, & Spruyt, 2012). In order to achieve comprehensive coverage of academic publications, the Flemish Government entrusted an independent body of academics with the task of selecting the peer reviewed outputs published since 2000 from the whole of the SSH publications submitted by the universities for inclusion in the VABB-SHW. This task proved especially difficult for book publications (Ghesquière, Van Bendegem, Gillis, Willems, & Cornelissen, 2011). After considerable debate, it was decided to include all the book publications by a limited number of 82 publishers that had been identified as the most prestigious and selective for the SSH in a similar exercise in Norway (Sivertsen, 2010; Engels et al., 2012). Three publishers were added in 2011 and another 33 in 2012. In the PRFS this latest selection of 118 publishers is applied to the preceding 10-year window, i.e. the period 2002-2011. In addition, in 2010 the Flemish publishers' association introduced the Guaranteed Peer Reviewed Content (GPRC) label in order to allow their members to make peer review of individual books explicit and to facilitate inclusion of these books in the VABB-SHW (Verleysen & Engels, 2013).

Data and methodology

The analysis presented in this paper is based on a dataset of 5140 book publications by academic scholars affiliated with a university in Flanders from the period 2002-2011. They belong to 16 SSH disciplines and two general categories. This total comprises 401 monographs, 762 edited books and 3.977 book chapters contained in the VABB-SHW. Of course, since the VABB-SHW collects SSH publications by scholars affiliated with a Flemish university, not all chapters that appeared in the 762 edited books are included. The majority of the book chapters did appear in edited books published by scholars not affiliated with a Flemish university. Hence, although some places of publication are included twice or more because an edited book as well as one or more chapter therein are included in the VABB-SHW, all places of publication are included in the analysis. Places of

publication were used as available in the VABB-SHW database. Whenever the data contained more than one place of publication the first one mentioned was used. Missing places of publication were searched for and added.
For the total of 5140 places of publication, the geographic coordinates (latitude and longitude) were added in decimal notation to their bibliographic description. Barycentres were determined for the 5 consecutive 2-year periods (2002-3: n=724, 2004-5: n=756, 2006-7: n=860, 2008-9: n=1248, and 2010-11: n=1389 [excluding places of publication of GPRC-labelled books for reasons of comparability]) by calculating the weighted average of the relevant coordinates. Weighting was done according to the number of publications for any given place of publication (Jin & Rousseau, 2001; Rousseau, 1989a; Rousseau, 1989b). The resulting barycentre coordinates were located on a Google map using the open software tool Geocommons.com.
One limitation of this approach is that places of publications may be very far apart (Rousseau, personal communication). This may complicate the interpretation of the result. Therefore a second set of barycentres was calculated for the 16 SSH disciplines, restricting the places of publication to European locations only. To this end Europe was defined as the EU (including its acceding or candidate members Croatia, Iceland, Montenegro, Serbia, the FYR of Macedonia and Turkey) plus Albania, Belarus, Moldavia, Norway and Switzerland. As Table 1 shows, a clear majority of 92% of the places of publications are European.

Table 1. Number of European and non-European places of publication per discipline.

Discipline	# of European locations	# of non-European locations	Total number of locations
Psychology (1)	79	38	117
Communication Studies (2)	92	16	108
Political Science (3)	251	45	296
Social Health Sciences (4)	26	6	32
Educational Studies (5)	161	28	189
Sociology (6)	141	19	160
Economics and Business (7)	325	35	360
Philosophy (8)	441	35	476
Art History (9)	225	17	242
History (10)	318	19	337
Literature (11)	605	34	639
Criminology (12)	104	5	109
Law (13)	436	26	462
Archeology (14)	60	2	62
Theology (15)	441	35	476
Linguistics (16)	660	18	678
All Disciplines	**4365**	**378**	**4743**

Results and discussion

The barycentre for all 5140 book publications from 2002-2011 is located at 50.08826° latitude and -2.69505° longitude. This corresponds to a location in the English Channel, situated some 50km to the South-West of Weymouth in Dorset, United Kingdom. This location, well outside Flanders or Belgium, demonstrates the importance of non-Flemish publishers for book publications by Flemish scholars. As the barycentre is located some 450 km to the West of Flanders, it is especially indicative of the weight of British and American publishers.

Comparing sub-periods

When comparing the barycentres of the five consecutive 2-year periods, it becomes apparent how the geographic centre of weight of scholarly book publishing has moved to the South-West marginally. The various locations again demonstrate the considerable and perhaps growing importance of British and North American publishers, especially in the most recent years.

(source: VABB-SHW; Geocommons.com; Google Maps)

Figure1. Barycentres for the periods A (2002-3); B (2004-5); C (2006-7); D (2008-9); E (2010-11), taking into account all publication places; and X (barycentre of the 5 Flemish universities' addresses).

Comparing disciplines

Barycentres for the 16 SSH disciplines lie at a varying geographic distance from the barycentre of the 5 Flemish universities (location X, figure 1). The locations of the 16 barycentres stretch out from a point near the Belgian-French border (Linguistics=16; Lat. 50.34762°, Long. 3.32916°) to the middle of the Atlantic Ocean between Brittany and Newfoundland (Psychology=1; Lat. 48.00301°, Long. -27.8229°). This alignment of the 16 barycentres along an East-West axis indicates considerable differences between disciplines regarding the importance of, respectively, Flemish, other continental European, British and American publishers.

Notable in Figure 2 is the discrepancy between the locations of the barycentres of the Social Science disciplines (SS) and those of the Humanities disciplines (H). The barycentres of all Humanities disciplines except Communication Studies are located in or near the Channel, while those of most Social Science disciplines are situated in the Atlantic Ocean to the South-West of Ireland, or, in the case of Psychology, even some 500km further to the West. This illustrates that American

publishers are of greater importance for the Social Sciences than for the Humanities.
When looking at the locations of the barycentres based on the European places of publication only, a marked contrast is noticeable between the Social Sciences and the Humanities as well.

(source: VABB-SHW; Geocommons.com; Google Maps)

Figure 2: Barycentres for 16 SSH disciplines (1=Psychology (SS); 2=Communication Studies (H); 3=Political Science (SS); 4=Social Health Sciences (SS); 5=Educational Studies (SS); 6=Sociology (SS); 7=Economics and Business (SS); 8=Philosophy (H); 9=Art History (H); 10=History (H); 11=Literature (H): 12=Criminology (SS); 13=Law (H); 14=Archeology (H); 15=Theology (H); 16=Linguistics (H)), taking into account all places of publication.

Figure 3. Barycentres for SSH disciplines (Europe) (1=Psychology (SS); 2=Communication Studies (H); 3=Political Science (SS); 4=Social Health Sciences (SS); 5=Educational Studies (SS); 6=Sociology (SS); 7=Economics and Business (SS); 8=Philosophy (H); 9=Art History (H); 10=History (H); 11=Literature (H): 12=Criminology (SS); 13=Law (H); 14=Archeology (H); 15=Theology (H); 16=Linguistics (H)), taking into account European places of publication only.

The Social Sciences' barycentres are mostly located in the United Kingdom or in the North Sea between the UK and the Netherlands, illustrating the weight of British publishers. The divergent result for Social Health Sciences (4) needs to be interpreted with caution, as data for this discipline is limited to 32 book publications (see Table 1). For all Humanties' disciplines but Law and Communication Studies, the barycentres are located within Flanders, pointing to the greater weight of Flemish and other continental European publishers.

Conclusion

The places-of-publication barycentre of monographs, edited books and book chapters published by scholars affiliated with Flemish universities in the period 2002-2011 is situated some 450 km to the West of Flanders and seems to be moving westwards slightly. This finding illustrates the importance of British and American publishers for SSH scholarly research at Flemish universities. At the same time, there is a marked difference between the Social Sciences and the Humanities. For the Social Sciences, British and American publishers are more important than for the Humanities. Humanities' researchers rely more on publishers situated in Flanders and elsewhere in continental Europe for the publication of their books and chapters.

The barycentre method thus proves to be very well applicable to book publications. As a first exploration of the method, and more in general of a geographic analysis of book publishing in the SSH, barycentres were located on an actual map of the world. In our presentation at ISSI, geometric representations of barycentres in standardised polygons (Rousseau, 2008) will be added in order to provide additional quantitative insight regarding the evolving role of continental European, British, American and other publishers over time. We conclude that the inclusion of book publications in the study of SSH publication patterns remains indispensable and that the barycentre method provides a useful addition for measuring aspects of research internationalisation.

Acknowledgments

We thank Ronald Rousseau for comments and suggestions during the preparation of this paper.

Reference List

Bartlett, A. A. (1985). U.S. population dynamics. *American Journal of Physics, 53,* 242-248.

Engels, T. C. E., Ossenblok, T. L. B., & Spruyt, E. H. J. (2012). Changing publication patterns in the social sciences and humanities, 2000-2009. *Scientometrics, 93,* 373-390.

Ghesquière, P., Van Bendegem, J.-P., Gillis, S., Willems, D., & Cornelissen, K. (2011). Het VABB-SHW: eerste versie klaar, nu verfijnen. In K.Debackere & R. Veugelers (Eds.), *Vlaams Indicatorenboek 2011* (pp. 260-264). Brussel: Expertisecentrum O&O Monitoring.

Hicks, D. (2004). The four literatures of social science. In H.F.Moed, W. Glänzel, & U. Schmoch (Eds.), *Handbook of quantitative Science and Technology Research: The use of publication and patent statistics in studies of S&T systems* (pp. 473-496). Dordrecht: Kluwer Academic.

Jin, B. & Rousseau, R. (2001). An introduction to the Barycentre method with an application to China's mean centre of publication. *Libri, 51,* 225-233.

Ossenblok, T. L. B., Engels, T. C. E., & Sivertsen, G. (2012). The representation of the social sciences and humanities in the Web of Science. A comparison of

publication patterns and incentive structures in Flanders and Norway (2005-9). *Research Evaluation, 21,* 280-290.
Rousseau, R. (1989a). Kinematical Statistics of scientific output. Part I: geographical approach. *Revue Française de bibliométrie, 4,* 50-64.
Rousseau, R. (1989b). Kinematical statistics of scientific output. Part II: standardized polygonal approach. *Revue Française de bibliométrie, 4,* 65-77.
Rousseau, R. (2008). Triad or Tetrad: another representation. *ISSI Newsletter, 4,* 5-7.
Sivertsen, G. (2010). A performance indicator based on complete data for the scientific publication output at research institutions. *ISSI Newsletter, 6,* 22-28.
Verleysen, F. T. & Engels, T. C. E. (2013). A Label for Peer-Reviewed Books. *Journal of the American Society for Information Science and Technology*, 64, 428-430. *Doi: 10.1002/asi.22836*.

MEASURING THE ACADEMIC IMPACT OF RESEARCHERS BY COMBINED CITATION AND COLLABORATION IMPACT

Jielan Ding [1], Liying Yang [1], Qing Liu [2]

[1] *dingjielan@mail.las.ac.cn, yangly@mail.las.ac.cn*
National Science Library of the Chinese Academy of Science
33 Beisihuan Xilu, Zhongguancun, Beijing, 100190 (P.R.China)

[2] *liuqing@mail.whlib.ac.cn*
The Wuhan Branch of the National Science Library of the Chinese Academy of Science
25 west of xiaohongshan, Wuchang Dist, Wuhan, 430071 (P.R.China).

Abstract
To evaluate the academic impact of researchers in a more comprehensive way, this paper utilizes both the citing and collaborating aspects which are the two main forms occurring in scientific communication. Three citation-based indicators and three collaboration-based indicators are selected and combined into two dimensions by using factor analysis. Then based on the position of the researchers in a two dimensional coordinate system, career roles and career paths of researchers can be revealed. Finally, we provide a theoretical framework for describing the career roles of researchers by the combination of citation impact and collaboration impact.

Introduction
In bibliometrics many studies focus on exploring the academic impact of researchers. It is well accepted that academic impact is produced during scientific communication; therefore measuring the academic impact should be based on the process of scientific communication. Citing and collaboration are the two main forms of scientific communication which can be measured by bibliometric methods. Many published results of investigations are related to citing and collaboration impact.

Most studies concentrated on citation analysis and the indicators used are citation-based, such as total citations, h-index, and citations per publication. This kind of studies stems from the impact in the process of citing, so we refer to this kind of research as "citation impact studies"(e.g. Garfield,1999; Hirsch,2005; Egghe,2006; Jin, Liang&Rousseau,2007; Moed, 2011; Leydesdorff &Bornmann, 2011; Rousseau, 2012).

Another way to detect the impact of researchers, different from using citation indicators is collaboration analysis. With the widely application of social network analysis (SNA) in the field of scientometrics, more and more researchers apply SNA measures to detect academic impact (e.g., Leydesdorff, 2007; Bollen, Van de Sompel, Hagberg &Chute, 2009), especially in collaboration network (e.g.,

Newman 2001a,b; Liu, Bollen, Nelson, & Van de Sompel, 2005; Rodriguez & Pepe, 2008). For researchers' impact study, the micro-level network indicators, such as degree centrality, closeness centrality, and betweenness centrality, are used for measuring researchers' impact in collaboration networks (Yan & Ding, 2009). This kind of research studies the impact produced in collaboration, so we refer to it as "collaboration impact studies".

There is no doubt that citation and collaboration can each describe one aspect of academic impact of researchers. To measure the academic impact of researchers in a more comprehensive way, the two aspects can be combined, in particular as they have a positive relationship (Yan & Ding, 2009; Levitt, Thelwall & Levitt, 2011). A researcher's citation impact is the degree of attention aroused by his academic achievement, so the citation impact represents his academic level to some extent. The collaboration impact reflects one's importance in a certain research community. So, we claim that a combined analysis reflects authors' status in a certain field.

This article aims to obtain both the citation impact and collaboration impact of researchers leading to a more comprehensive way of measuring academic impact. Using a set of indicators based on the two dimensions and factor analysis, we attempt to describe: (1) the career roles of researchers by a combined analysis of the two dimensions; (2) the career paths of researchers by measuring changes in both dimensions.

Methodology

Datasets

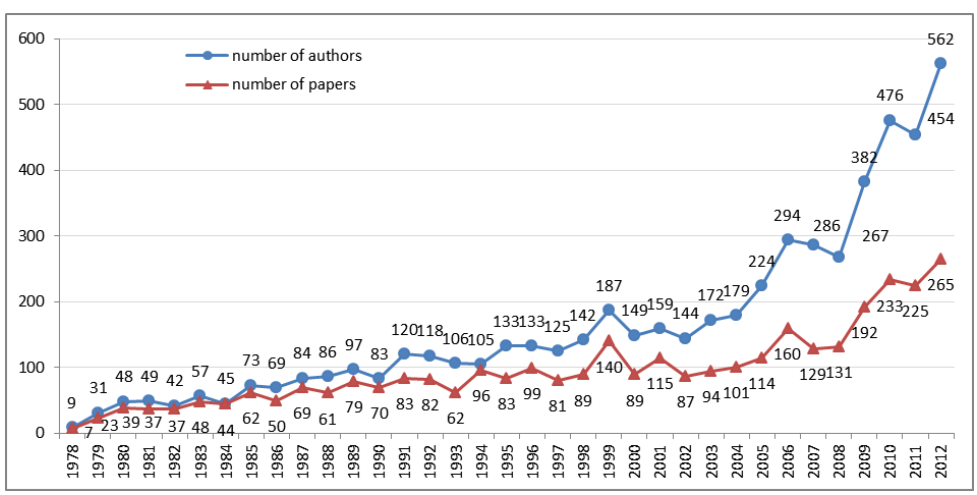


Figure 1. The number of authors and papers of *SCIENTOMETRICS* in 1978-2012. (download date:2013.01.08)

We took the field of scientometrics as an example to apply and test our method. The study focuses on the combined method, and the journal *SCIENTOMETRICS* is the most specialized and typical journal in this field, so the study selected all the papers from *SCIENTOMETRICS* as dataset, though one journal on its own can never describe the entire career of a scientist.

We download the data of *SCIENTOMETRICS* from the Web of Science on 8 January, 2013. The time span is from 1978 to 2012. There are 3376 papers and 3419 different authors after manual data cleaning of authors in the dataset. The yearly data of number of authors and papers of *SCIENTOMETRICS* is shown in Figure 1. Because the increase in the number of papers and authors after the year 2000 is bigger than that before the year 2000, we chose the year 2000 as the cut-off point for detecting the change of impact in the two periods which is the reference year for describing the career paths of the authors. The distribution of authors is shown in Table 1, and the data shows that about 10% authors, who published at least four papers in the period 1987-2012 published over 50% of all papers.

Table1. The distribution of authors.

threshold of authors	*1978-2012 (total dataset)*				*1978-1999 (sub-dataset I)*				*2000-2012 (sub-dataset II)*			
	authors		*papers*		*authors*		*papers*		*authors*		*papers*	
	counts	*%*	*counts*	*%*	*counts*	*%*	*counts*	*%*	*counts*	*%*	*counts*	*%*
published at least 5 papers	223	6.5%	1797	53.2%	81	6.9%	677	47.0%	134	5.4%	911	47.1%
published at least 4 papers	325	9.5%	2018	59.8%	119	10.2%	785	54.5%	210	8.5%	1065	55.0%
published at least 3 papers	495	14.5%	2271	67.3%	175	14.9%	694	48.2%	340	13.8%	1245	64.3%
published at least 2 papers	982	28.7%	2663	78.9%	342	29.2%	1088	75.5%	688	27.9%	1475	76.2%
total	3419	100.0%	3376	100%	1172	100%	1441	100%	2470	100%	1935	100%

Indicators for the two dimensions

A single indicator describes just one aspect of academic impact, but impact is multi-dimensional. Therefore for each dimension we chose a group of indicators from different perspectives to measure the academic impact of researchers.

For the citation impact dimension, total citations, CPP, and h-index were chosen to measure the researchers' citation impact as these are commonly used for academic impact evaluation. We calculated the three indicators based on the citations which the papers published in *SCIENTOMETRICS* received from all papers in WOS.

For collaboration impact dimension, closeness centrality, which means one divided by the total geodesic distance from a node to all others, was chosen to measure the authors' impact over the entire collaboration network, and degree centrality, which means the number of neighbors of a node, was chosen to measure the authors' impact over the local collaboration network. The

collaboration ratio, which means the ratio of collaboration papers among all papers one published, was chosen to measure the depth of collaboration impact. Closeness centrality and degree centrality are network indicators which were calculated in PAJEK based on the *SCIENTOMETRICS* -collaboration network.
We calculated the selected six indicators for each author in the total dataset. These indicators form the base for measuring their career roles. Career paths are based on these six indicators, calculated separately for the sub-dataset I (1978-1999) and the sub-dataset II (2000-2012).

Factor Analysis

Table 2. Rotated Component Matrix[a].

The total dataset of 1978-2012 (KMO: 0.674)			*The sub-dataset of 1980-1999 (KMO: 0.659)*			*The sub-dataset of 2000-2012 (KMO:0.679)*		
	Component			Component			Component	
	1	2		1	2		1	2
total_cites	.927	.135	total_cites	.963	.149	total_cites	.928	.104
CPP	.628	.033	CPP	.544	.171	CPP	.748	-.047
h_index	.936	.166	h_index	.928	.193	h_index	.925	.119
collaboration ratio	-.205	.884	collaboration ratio	-.022	.906	collaboration ratio	-.180	.877
degree centrality	.520	.642	degree centrality	.392	.750	degree centrality	.490	.680
closeness centrality	.359	.622	closeness centrality	.332	.764	closeness centrality	.518	.422

Extraction Method: Principal Component Analysis. Rotation Method: Varimax with Kaiser Normalization. a. Rotation converged in 3 iterations.

We chose factor analysis to reduce the six indicators to two integral indicators in order to measure authors' citation and collaboration impact. We did this for two reasons: on the one hand, factor analysis is an objective weighted method of detecting the relationship among a group of indicators; on the other hand, the three citation-based indicators and the three collaboration-based indicators have a positive relationship and their interpretation overlaps to some extent. Therefore these indicators can't be simply added.
For the authors who published at least four papers in the total dataset, we took factor analysis based on their six indicator scores to obtain their citation impact and collaboration impact to describe their career roles. We did the same thing to the authors who published at least four papers in the sub-dataset I (1978-1999) and in the sub-dataset II (2000-2012) in order to detect their career paths.
After factor analysis by SPSS 16.0, the six indicators were reduced to two main components for the total dataset and the two sub-datasets. The results are shown in table 2. A varimax rotation was applied to measure loadings in order to make the components easier to interpret. For the three datasets (sub-datasets), most loadings of the three citation-based indicators are on the first component which we name "citation impact" and most loadings of three collaboration-based

indicators are on the second component which we refer to as "collaboration impact". We note that the two components are orthogonal so there is no linear dependence between the two.

Measuring career roles and career paths for each author

After factor analysis, taking the citation impact component as the abscissa and the collaboration impact component as the ordinate axis, we got each author's location in the Cartesian coordinate system. Each author is located in one of the four quadrants in the two dimensional Cartesian coordinate system. As the four quadrants represent four different career roles we can describe the career role of each author.

To some extent, the ascent and descent of a researcher's citation impact could represent the change of his academic level while the change of a researcher's collaboration impact could represent the change of his activity in collaboration. So the change of authors' impact on the two dimensions can reveal the career paths of researchers.

Results

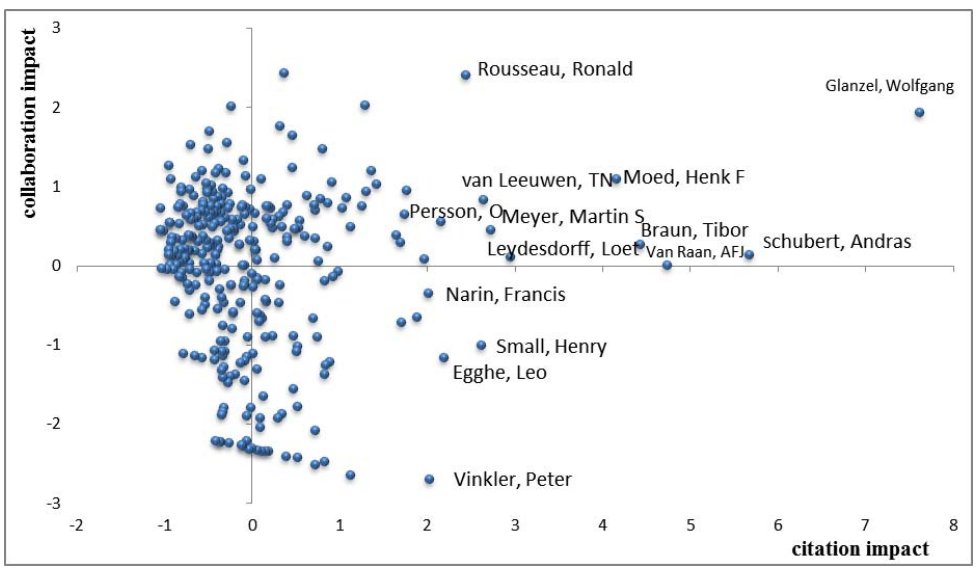
The career roles of researchers in SCIENTOMETRICS

We took the 325 authors in *Scientometrics* who published at least four papers in the period 1978-2012 as examples to describe their career roles by using the two dimensional coordinate system. Then we got the location of each author in the four quadrants which represent four different career roles. The results are shown in figure 2 and the extraordinarily high citation impact authors whose citation impact scores are higher than two are shown on the map. There are 64 authors located in the top right quadrant, 50 authors located in the bottom right quadrant, 139 authors located in the top left quadrant and 72 authors located in the bottom left quadrant.

The authors located in the top right quadrants have both high citation impact and collaboration impact. They not only published highly cited papers which appeal to a lot of people, but also play an important role in scientific collaboration. These authors usually have a very prestige in the field of scientometrics, for example "Glänzel, Wolfgang", "Schubert, Andras", "Van Raan, AFJ", "Braun,Tibor", "Moed, Henk F", "Leydesdorff, Loet", "Meyer, Martin S", "van Leeuwen, Thed N", "Rousseau, Ronald", "Persson, Olle", etc. These high citation and collaboration impact authors are the excellent and core authors.

The authors located in the bottom right quadrant have high cited papers but they rarely collaborate or they are located in the periphery of the collaboration community. Taking "Small, Henry", "Egghe, Leo", "Vinkler, Peter" for example, they are excellent researchers in scientometrics who are Derek John de Solla Price award winners, but they seldom collaborate, especially "Vinkler, Peter" who never collaborated in all his 31

SCIENTOMETRICS papers. These authors of high citation impact but low collaboration are the excellent and lonely (like to work alone) authors.
The authors located in the top left quadrant are very active in collaboration, but relatively ordinary in academic impact. Most internationally oriented Chinese authors who have high prestige in China in the field of scientometrics are placed in this quadrant. Authors of high collaboration impact but low citation impact are the ordinary and core authors.
The authors located in the bottom left quadrant have both low citation and collaboration impact who are the ordinary and lonely authors. They are ordinary in academic achievement and in the periphery of the collaboration network.



(The origin of coordinates is (0,0) which means the average level of the citation and collaboration impact. The authors whose citation impact score are higher than 2 are marked with their names)

Figure 2. The career roles of the 325 authors in *SCIENTOMETRICS* who published no less than 4 papers in year 1978-2012.

The career paths of researchers in SCIENTOMETRICS

We defined seven career paths according to the change of position based on impact score in two periods for each dimension. These are shown in Table 3. There are seven career paths for each dimension; therefore there are 49 combinations representing 49 different possibilities of authors' career status shown in table 4. highly characteristic highly characteristic
We focused on the career path of top authors, and we took the top 20% authors as the top authors for the sub-dataset of year 1978-1999 (first period) and the sub-dataset of year 2000-2012 (second period) for each dimension. There are 96 "top authors" who get into the list of the top authors in citation or collaboration

dimension in at least one period. They form the examples for our career paths analysis. Their career paths are shown in Table 5.

Table 3. The definition of seven career paths.

Description of authors	*career paths*	*the score in first period*	*the score in second period*
top authors	Plateau	>=A	>=B
	New force	—	>=B
	Fall	>=A	—
	Rise	< A	>=B
	Decline	>=A	< B
others	Go up	< A	< B, and higher than the score of former period
	Go down	< A	< B, and lower than the score of former period

(A and B are thresholds for selecting the top 20% authors for each time period which are shown in table 4. "—" means the author didn't have a score because he published less than 4 papers and didn't occur in sample dataset.)

Table 4. Values of A and B for selecting the top 20% authors.

dimensions	*A (threshold for first period)*	*B (threshold for second period)*
citation impact	0.41	0.47
collaboration impact	0.95	0.63

Table 5. The career paths of the 96 top authors (profile).

collaboration impact / *citation impact*	*plateau*	*new force*	*rise*	*go up*	*go down*	*decline*	*fall*
plateau	-	-	2	3	6	4	-
new force	-	6	-	15	-	-	-
rise	-	-	1	2	2	2	-
go up	3	27	2	-	-	1	-
go down	-	-	-	-	-	1	11
decline	-	-	-	1	1	1	-
fall	-	-	-	-	5	-	1

The authors whose career path is "new force" in citation dimension and "going up" in collaboration dimension are likely the "new force" who shifted from other fields to this field. They are absent in citation impact dimension in this field in the first period, but jump to be top researcher by publishing some highly cited papers in the second period. At the same time they have not yet constructed a broad partnership in this field. Porter, Alan L. is such an example (being an expert in data mining and industry-university relations).

The authors whose career path is “new force” both in citation dimension and collaboration dimension are likely the new excellent researchers who grow up in an excellent community of this field. They didn’t occur in the first period dataset, but jumped to be a top researcher in both dimensions in the second period. Liang, Liming is a case in point, collaboration with Rousseau, Ronald.
The authors whose career path is “plateau” in the collaboration dimension and “going up” or “rising” in the citation dimension are located in the center of the collaboration network work in the two periods and become more excellent in academic achievement in the later period, such as “Noyons, ECM”.
The authors whose career path is “plateau” in citation dimension and “rising” or “going up” in collaboration dimension are excellent in academic achievement and become active in collaboration and get to the core position of the collaboration network. Taking “Glänzel, Wolfgang”, “Rousseau, Ronald”, “Leydesdorff, Loet” for example, their citation impact was very high in the two periods and in the second period they get to be located in the central position of the collaboration network.
Most internationally oriented Chinese authors’ career path is “new force” in collaboration dimension and “going up” in citation dimension. Collaboration obviously increases their impact.
The authors who retired and stopped doing research in scientometrics end up in a “going down” or “falling” position. Taking “Narin, Francis” for example, his career path is “falling” in citation dimension and “going down” in collaboration dimension. He established CHI in 1968, an internationally recognized research consultancy company specializing in developing evaluation tools and indicators for science and technology analysis, and obtained the Derek John de Solla Price award in 1988. He retired from CHI in 2004.

Conclusion

Since scientific collaboration becomes more and more popular, it is well accepted that researchers’ citation impact and collaboration impact are equally essential. This investigation took the citing and collaboration dimensions simultaneously into account, leading to a new approach to measure the academic impact of researchers in a more comprehensive way.
By combined analysis of citation impact and collaboration impact, we can discover the detailed information about authors’ career status, such as career roles or career paths. Based on empirical results, we provided a framework to describing the career roles of researchers which is shown in Figure 3. The origin of coordinates means the average level. A researcher’s citation impact represents his academic level. We used “excellent” to describe researchers when their citation impact is higher than the average level and used “ordinary” to describe researchers when their citation impact is lower than the average level. The collaboration impact can usually reflect one’s importance in a research community. We used “core” to describe researchers when their collaboration impact is higher than the average level and used “lonely”(like to work alone) to

describe when their collaboration impact is lower than the average level. Thus we can describe the researchers by using four career roles. If someone has a high score in both citation impact and collaboration impact, he/she is probably an excellent and core researcher. If someone has a high score in citation impact but a low score in collaboration impact, that person is very likely an excellent and lonely researcher in the field. If someone has a high score in collaboration impact but low score in citation impact, he is probably an ordinary and core researcher. If someone has a low score in both citation impact and collaboration impact, he is just an ordinary and lonely researcher (at least at the moment of investigation). For S&T policy makers, identifying researchers in the quadrant we proposed may help them in finding key researchers, possibly with high collaboration (social) skills.

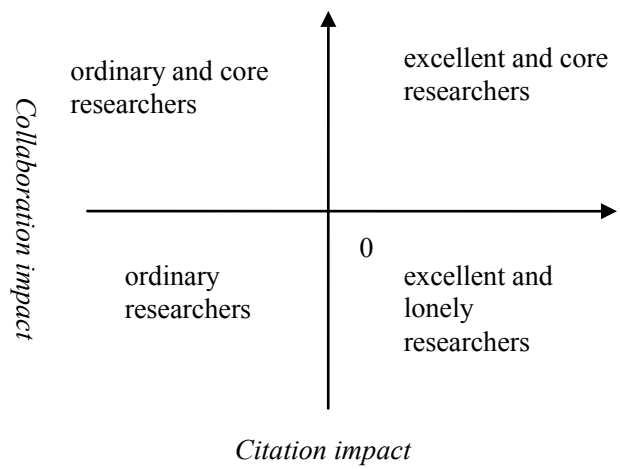


Figure 3: A framework for revealing the career roles of researchers

By detecting the impact changes in two dimensions in different time windows, we could distinguish various career paths for each researcher. The researchers who are top researchers in citation impact for the two periods and become more active in collaboration impact are the "evergreen tree" scientists of this field. The "new force" researchers are those who were not main researchers (they published less than 4 papers) in the first period and turned to be top researchers in the second one. They could perhaps be described as "dark horses" in this field. Finding those different types of researchers is also a way to evaluate researchers from different perspectives.

In empirical study, we detected the career roles and career paths of authors in *SCIENTOMETRICS*. The result may not reflect the real lifetime career roles or paths for researchers as the dataset is limited to one journal, but it certainly provides (partial) information of scientists' profiles active in our field. We use the

combined method in the field of scientometrics and we will do further study of applying the method in other fields to detect the validity of the method.

Acknowledgements

The study was supported by Knowledge Innovation Program of The Chinese Academy of Sciences (Project NO. 10XZNL9). We thank Ronald Rousseau for giving critically valuable advice and making a lot of linguistic corrections. We further thank Prof. Alan Porter (Georgia Tech) and Ting Yue (National Science Library of CAS) for helpful discussions.

References

Garfield,E. (1999). Journal impact factor: A brief review. Canadian Medical Association Journal, 161(8),979–980.

Hirsch, J. E.(2005). An index to quantify an individual's scientific research output. Proceedings of the National Academy of Sciences of the USA,102 (46) : 16569- 16572.

Egghe, L.(2006). Theory and Practise of the g-index. Scientometrics, 69(1):131-152.

Jin, B.H.; Liang, L.M. &Rousseau, R.(2007). The R- and AR-indices: Complementing the h-index. CHINESE SCIENCE BULLETIN, 52(6): 855-863.

Moed, H.F.(2011). Measuring contextual citation impact of scientific journals. JOURNAL OF INFORMETRICS,4(3): 265-277.

Leydesdorff,L. &Bornmann,L.(2011). Integrated Impact Indicators Compared With Impact Factors: An Alternative Research Design With Policy Implications. JOURNAL OF THE AMERICAN SOCIETY FOR INFORMATION SCIENCE AND TECHNOLOGY,62(11): 2133-2146.

Rousseau, R.(2012). Basic Properties of Both Percentile Rank Scores and the I3 Indicator. JOURNAL OF THE AMERICAN SOCIETY FOR INFORMATION SCIENCE AND TECHNOLOGY,63(2): 416-420.

Bollen, J. , Van de Sompel, H. , Hagberg, A. & Chute, R. (2009). A principal component analysis of 39 scientific impact measures. *PLoS ONE* ,4(6), e6022.

Levitt, M.J., Thelwall,M., & Levitt, M. (2011). To what extent does the citation advantage of collaboration depend on the citation counting systems? In: (E. Noyons, P. Ngulube & J. Leta, Eds.) *Proceedings of ISSI 2011—13th International Conference of the International Society for Scientometrics and Informetrics, pp.*398–408. Durban: ISSI, Leiden University and University of Zululand.

Leydesdorff, L. (2007)."Betweenness Centrality" as an Indicator of the "Interdisciplinarity" of Scientific Journals. *Journal of the American Society for Information Science and Technology*, 58(9), 1303–1319.

Liu, X., Bollen, J., Nelson, M.L.,& Van de Sompel, H. (2005). Co-authorship networks in the digital library research community. *Information Processing and Management*, 41,1462-1480.

Newman, M.E.J. (2001a). Scientific collaboration networks: I. Network construction and fundamental results. *Physical Review E*, 64, 016131.
Newman, M.E.J. (2001b). The structure of scientific collaboration networks. *Proceedings of the National Academy of Science of the United States of America*, 98(2), 404-409.
Rodriguez, M.A. & Pepe, A. (2008). On the relationship between the structural and socioacademic communities of a coauthorship network. *Journal of Informetrics,* 2(3), 195-201.
Yan, E.J. & Ding, Y. (2009). Applying centrality measures to impact analysis: a coauthorship network analysis. *Journal of the American Society for Information Science and Technology*, 10, 2107-2018.

MEASURING THE EXTENT TO WHICH A RESEARCH DOMAIN IS SELF-CONTAINED

Alan L. Porter[1], David J. Schoeneck[2], Stephen J. Carley[3]

[1]*alan.porter@isye.gatech.edu*
Technology Policy & Assessment Center, Georgia Tech, Atlanta GA, USA 30332-0345,
and
Search Technology, Inc., Norcross GA USA 30092

[2]*daves@searchtech.com*
Search Technology, Inc., Norcross GA USA 30092

[3]*stephen.carley@innovate.gatech.edu*
Program in Science, Technology & Innovation Policy, Georgia Tech, Atlanta GA, USA 30332-0345

Abstract
For several years we have been working to measure cross-disciplinarity, especially trying to determine interdisciplinary integration of diverse knowledge. Existing indicators, particularly our own Integration and Diffusion scores, speak to disciplinary engagement, but not directly to whether knowledge is being transferred from areas heretofore not well-connected to a research domain. This paper introduces simple metrics that gauge 1) the extent to which a research domain references papers generated within that domain vs. outside publications, and 2) the extent to which a domain's publications are cited within itself vs. outside. We address three emerging technologies as case research domains – Nano-Enabled Drug Delivery, Hybrid & Electric Vehicles, and Dye-Sensitized Solar Cells. We first tabulate and map their disciplinary locations based on Web of Science Categories. We then calculate the new metrics to offer additional perspectives on knowledge diffusion.

Conference Topic
Scientometrics and Indicators – new developments (Topic 1)

Introduction
Measuring interdisciplinarity is of interest in terms of research planning and assessment, and social studies of science. Stirling (2007) presents a compelling conceptual basis for considering diversity in terms of variety, balance, and disparity. This fits well with an indicator for the "integration" of disparate knowledge (c.f., Porter et al., 2008) and its application (c.f., Porter and Rafols, 2009). For a comprehensive treatment of measuring interdisciplinarity, see the review of the state of the art by Wagner and colleagues (2011). The Integration score just noted gauges the diversity of the references cited by a given paper or set of papers. Conversely, one can address the diversity of document sets that cite a

paper or set of papers (Zitt & Small, 2008). In this respect, Carley and Porter (2012) provide an index to measure forward diversity – called a Diffusion score.

Many such indicators make use of existing categorizations, especially the Web of Science (WoS) Subject Categories (WOSCs)[112] that are treated here. Rafols and Meyer (2010) point out that diversity in cited WOSCs does not assure true integration of disparate knowledge sources in a given body of research. Our measures rely on Web of Science categories, and do a reasonable job of depicting disciplinary engagement. However, they don't really address whether publications and citations reflect knowledge transfer among disparate areas, or just that a research domain straddles multiple WOSCs

Interest in assessing the extent to which a given research area draws upon "outside" knowledge prompts the present inquiry. We address three research domains and inquire to what extent they cite research not inherently part of the target research domain, and to what extent the research they produce interests others (i.e., is cited outside the domain).

Over the past few years, several of us have been analyzing Dye-Sensitized Solar Cells (DSSCs) – an emerging photovoltaic technology. We have observed that this research domain seems remarkably cohesive. For instance, in a set of 4,104 records from WoS through 2010, without cleaning to consolidate variations on citations of the same paper, we find one paper cited 2,396 times (i.e., by 58% of the DSSC papers in the whole set). Moreover, 101 papers are cited 100 or more times, by these 4,104 papers. Social network maps of co-authors, co-cited authors, or such show strong interconnection.

In the past months, we have been studying two other R&D domains – Hybrid & Electric Vehicles (HEVs) and Nano-Enabled Drug Delivery (NEDD). HEVs reflect a more mature technology with multiple sub-systems. NEDD pulls together materials science with bio-medical research, incorporating a multitude of intertwined drug, transport mechanism, and target possibilities. Casual discussion surfaced our sense that these research domains are much less cohesive than DSSCs. As a simple indicator, among 61,465 NEDD records from WoS through 2012, the most cited reference (before consolidating) has only 1,538 hits (so only 2.5% of the retrieved NEDD papers cite any one reference – in contrast to the 58% just noted for DSSCs).

[112] One needs to be aware that since late 2011WoS provides this information in the field called "WC" that formerly was called "SC"; they provide other, somewhat more aggregated categories under the label, "SC." This paper consolidates old SC and new WC information -- there are some 8 categories unique to the old SCs and some 7 unique to the new WCs, of the set of some 224 covering Science Citation Index and Social Sciences Citation Index WOSCs, plus additional ones for the Arts and Humanities Citation Index.

Reflecting on the apparent differences among these three research domains sparks two research questions that this paper addresses:

1. Q1: For a given research domain, to what extent is knowledge internal, in the sense that papers cite references within the domain (vs. external papers – i.e., those not retrieved by the domain search)?
2. Q2: Conversely, to what extent is the knowledge generated in that domain of external interest, in the sense that papers published receive relatively high proportion of cites (to them) from outside?

This paper sets forth to measure citation patterns for the three research domains noted – DSSCs, HEVs, and NEDD. We believe results will be of interest in our studies of these emerging science & technology areas (these are completely separate analyses). We are also interested in the potential of such citation metrics to help in the study of research knowledge diffusion patterns and the forces that promote or impede transfer processes. Assessing how "porous" research domains are, in terms of citation in and out, would seem to offer potential insight into how self-contained research areas are. We are not aware of others having done this. Of course, citations are made for multiple motives and have their limitations in indicating knowledge transfer at work.

Data and Methods

Various colleagues have worked out suitable search strategies for the three datasets under study here (see Appendix). Using those, search sets of WoS full records, including Cited References (CRs) have been downloaded recently. The date ranges vary somewhat so that is a consideration (e.g., might citation patterns be changing over time?). The basic data include (see Appendix for details):

- DSSCs – 8919 records from 1991 through 2012 (peaking at 1924 records in 2012; note that 2012 publications are not completely indexed by WoS at the time of search)
- HEVs – 7323 records from 2000 through 2012 (peaking in 2011,with 2012 incompletely indexed at 1025 records)
- NEDD – 61,465 records from 2000 through 2012 (with 14 in 2013, peaking with 9463 in 2011, and 2012 incompletely indexed, with 8120 records downloaded).

In each of these WoS abstract record sets, there are multiple document types, dominated by journal articles and proceedings papers. The frequency of citation is highly skewed, as one would expect, and the overall number of Cited References (CRs) is large. For instance, for NEDD:

- 1,141,623 CRs are identified with 1 or more cites by the 61,465 papers
- 366,890 CRs with 2 or more cites
- 42,735 CRs with 10 or more cites
- 1045 CRs with 100 or more
- 19 CRs with 500 or more

We apply thesauri in VantagePoint [www.theVantagePoint.com], software designed to facilitate analyses of field-structured records, especially R&D publication and patent sets such as these. These thesauri help provide information on:

- Cited WOSCs, based on extraction of Cited Journal information from the CRs, then application of a Find & Replace thesaurus to help standardize nomenclature, followed by a thesaurus that links those journals to WOSCs
- Macro-Disciplines and Cited Macro-Disciplines, based on analyses of WoS journal-to-journal cross-citation, converted to WOSC-to-WOSC cross-citation, then factor analysed to group the WOSCs based on their affinities (see Leydesdorff and Rafols, 2009; Rafols et al., 2010; Leydesdorff et al., 2013). This WOSC citing matrix is also essential for the science overlay mapping coming up.

For this study, we encountered a decision about which CRs to examine (i.e., we could not analyze all of them). Table 1 compares alternative selection possibilities. The "Top 200" seem to offer a reasonable middle ground for our purposes. We expand slightly to take ties – e.g., 212 HEV CRs had 20 or more cites. [We note that these are uncleaned counts.]

Table 1. Highly Cited References

Criterion	DSSCs	HEVs	NEDD
Research Domain Set Size (# of papers)	8,919	7,323	61,465
100 or more cites by this many papers in the research domain set	310	10	1,045
Top 200 Cited References are cited at least this many times	134	20	222

We next take these target ~200 CRs for each research domain and apply fuzzy matching routines to identify additional CRs that likely refer to the same papers. These include two important variations: a) CRs with and without DOI information; and b) CRs with variant information (e.g., the seminal O'Regan and Graetzel DSSC paper, 1991, cited 5,192 times by these 8,919 papers, we find frequent variants -- e.g., O'Reagan or slightly variant page and volume information).

For these top cited ~200 in each dataset, we filtered the CR field text to take first author's name, year, and first journal title word (after testing some alternatives). We then made a group in VantagePoint of ~650 CRs, in the NEDD set that match those (similar process used for HEVs & DSSCs). We make that into a thesaurus and apply it to consolidate these ~200 CRs. This concentrates the CRs – e.g., the most cited one for NEDD increases from 1,538 to 1,576. This process is not perfect, but satisfactory – hand-checking indicates 1,578 (adding one cited only by last name, not adding another also published in *Nature* in 2001 with a different

page number). The second most cited CR increases from 1,279 to 1,304; hand-check suggests 1,302 (leaving out two in different issues of *Nature*, same year). The third most cited jumps from 1,090 to 1,352 (hand-check agrees). So, this consolidation helps. After consolidation, further analyses key on these highly cited references:

- 190 NEDD CRs, 207 HEV CRs, and 198 DSSC CRs.

Our design to address the two research questions uses the highly cited CRs to do "double-duty." To address Q1 (looking back to see how much of the cited knowledge is internal), we check to see how many of these CR papers are themselves part of the paper set. We limit this inquiry to high CRs published in the time frame of the research domain set (i.e., 2000-2012 for NEDD and HEVs; 1991-2012 for DSSCs).

To address Q2, we study the subset of those high CRs that are included in the research domain set (just identified in addressing Q1). We search for those records to capture their current Times Cited in WoS ("TC"), then compare those values to the # of times each is cited by the respective research domain papers.

For the 207 HEV high CRs, we located and downloaded 175 (31 of the 32 not found had no DOI, so apt to be items not indexed by WoS – e.g., books). Of the 175, 144 were published 2000-2012 (in the time range covered by our HEV paper set). Combining information from the HEV papers file with that from the downloaded CR records, we check TC against HEV TC (cites by the HEV papers set) for those 144 papers. In 7 of 144 cases, this shows negative, and in 2 it is zero, so we recheck and correct. We find many CRs that are hard to classify. In checking these 9, for instance, we have to deal with multiple CRs for that author, that year, and the same first word in the source title. We also find CRs with differences on some of those fields, but compelling matches on other fields – DOI, author, year, volume, and/or page #. We sometimes see differences in author initials; we see some with a term in front of the typical cited source; and so forth. Some CRs without DOI or page # are a judgment call (we manually do so by taking into account what other papers show by that first author in the CRs). After checking, 3 HEV high CRs still show as negative on our best estimates – 1 shows 1 more citation by the HEV papers than in all of WOS; 2 show 33 more. For purposes of this analysis, we set all of those to be the same (i.e., TC = HEV TC) – i.e., all observed cites coming from the HEV records set.

We don't go into such detail for the other two sets. For NEDD, consolidation of the ~200 reduced to 188, of which 148 were published in the period, 2000-2012; of those 96 are in the NEDD papers set. For DSSCs, consolidation of the ~200 reduced to 198, of which 193 were found in WoS. We set aside 3 published prior to 1991, leaving 190 for further analyses.

Results

We first present descriptive statistics and disciplinary maps to gain a sense of what these three research domains encompass.

We locate the respective bodies of research on science overlay maps (Rafols et al., 2010) in Figure 1. These are scaled unequally (because the NEDD dataset is so much larger) just to show the relative disciplinary concentrations. The lower right map is the base map reflecting all WoS publications in 2010, with the nodes indicating 224 WOSCs and the labels showing the 19 Macro-Disciplines. The three research area maps overlay the respective publication intensity (larger node size indicating more publications in journals associated with that WOSC).

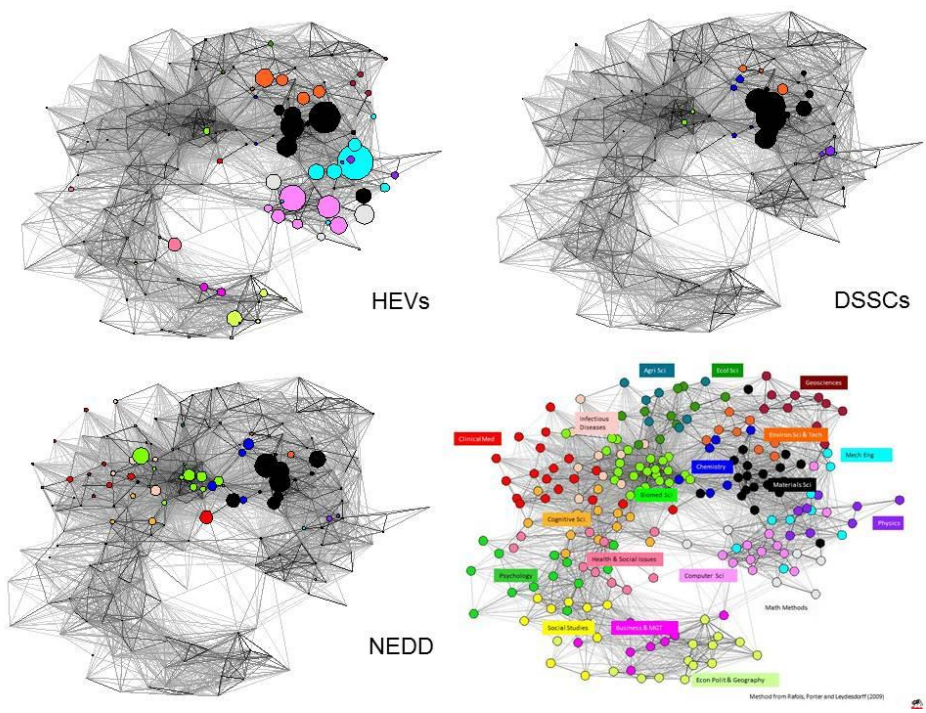


Figure 1. Overlaying the 3 Research Domains over a Base Map of Science

- HEVs show a highly multidisciplinary picture, led by Computer Science areas, with notable contributions in Materials Sciences, Environmental Science & Technology, and Mechanical Engineering. Note that this research domain, unlike the others, has considerable activity in Mathematical Methods, and in Economics, Political Science & Geography.
- DSSC research concentrates heavily in Materials Sciences, with notable contributions in Environmental Science & Technology, followed by Chemistry and Physics.

- NEDD shows major activity in Biomedical Sciences and Materials Sciences, followed by notable research publication in Clinical Medicine, Chemistry, and Infectious Diseases. There is also extensive work appearing in Cognitive Sciences, Environmental Science & Technology, and Physics.

We also tabulate the relative frequencies for publication Macro-Disciplines (MDs) (using record counts) and Cited Macro-Disciplines (using instances – i.e., if a paper cites 20 Physics papers, we tally those as 20 rather than just counting this as 1 record citing Physics). We then divide by the totals for each dataset and take percentages of the totals. Table 2 compares the 19 MDs and Cited MDs for each research domain. These provide MD values that correspond to the research concentrations visualized in Figure 1 in the "MDs (pubs)" columns. The prevalence of Materials Sciences is apparent.

Table 2. Research Domain Publication & Citation Concentrations by Macro-Disciplines

	HEVs		DSSCs		NEDD	
Macro-Disciplines	**MDs (pubs)**	**Cited MDs**	**MDs (pubs)**	**Cited MDs**	**MDs (pubs)**	**Cited MDs**
	% records	% instances	% records	% instances	% records	% instances
Agri Sci	0.1%	0.1%	0.1%	0.2%	0.8%	1.4%
Biomed Sci	0.5%	1.7%	1.8%	4.6%	40.9%	48.6%
Business & MGT	0.8%	1.3%	0.0%	0.0%	0.0%	0.0%
Chemistry	0.2%	0.7%	6.5%	5.5%	6.3%	4.5%
Clinical Med	0.2%	0.2%	0.2%	0.2%	8.8%	7.4%
Cognitive Sci	0.1%	0.4%	0.1%	0.1%	2.8%	2.7%
Computer Sci	41.4%	29.5%	2.3%	0.3%	0.3%	0.2%
Ecol Sci	0.3%	0.3%	0.1%	0.1%	0.2%	0.1%
Econ Polit & Geog	2.2%	2.4%	0.0%	0.0%	0.0%	0.0%
Environ Sci & Tech	16.3%	17.7%	12.6%	6.7%	1.3%	0.9%
Geosciences	0.9%	0.6%	0.1%	0.1%	0.1%	0.1%
Health & Social Issues	1.4%	1.1%	0.0%	0.0%	0.2%	0.2%
Infectious Diseases	0.0%	0.1%	0.1%	0.1%	4.4%	4.9%
Materials Sci	17.1%	32.8%	71.1%	78.8%	32.1%	27.2%
Math Methods	4.9%	2.6%	0.4%	1.4%	0.2%	0.7%
Mech Eng	12.3%	7.4%	0.8%	0.6%	0.4%	0.4%
Physics	1.1%	1.1%	3.9%	1.2%	1.2%	0.6%
Psychology	0.1%	0.0%	0.0%	0.0%	0.0%	0.0%
Social Studies	0.1%	0.1%	0.0%	0.0%	0.0%	0.0%

However, the main interest is to get a coarse view of the degree to which these research domains draw upon knowledge concentrated in the same, or other, disciplines in which they publish. Some observations:

- HEVs: Publish considerably more heavily in, than they cite, journals that aggregate to the Computer Sciences, Mechanical Engineering, and Math

Methods. Conversely, they draw more heavily than they publish in Materials Sciences and several less central MDs (e.g., Biomedical Sciences, Chemistry). This suggests research knowledge diffusion from the sciences toward engineering applications.

- DSSCs: Publish relatively more than they cite Environmental Science & Technology, Physics, and Computer Sciences; cite relatively more heavily Biomedical Science and Math Methods. This is a different distribution than the HEVs show, but also suggests knowledge transfer from fundamental science and math toward application considerations (e.g.,environmental).
- NEDD: Publications generally track citations.
- The above notes point to interesting differences, but also note the dominant tendency to publish where you cite – e.g., DSSCs in Materials Sciences; NEDD in Biomedical Sciences and Materials Sciences.

To address research question Q1, we investigate the "~200" heavily cited references of each dataset, leaving out those published prior to our search year range. We check whether those CRs are themselves included in the dataset. Results show:

- NEDD: 96 of 148 = **65% "internal"** (i.e., high CRs themselves in the NEDD papers set)
  and in number of cites by the NEDD paper set to these 148 papers – 34,982 are to those 96 internal papers and 18,807 are to the other 54 "external" papers = 65% internal also
- HEV: 114 of 144 = **79% internal**
  and in number of cites by the HEV paper set to those 144 papers – 5,062 are to those 114 internal papers and 885 are to the other 30 external papers = 85% internal
- DSSCs: 172 of 190 = **91% internal**
  and in number of cites by the DSSC paper set to the 190 papers – 50,470 are to the 172 internal papers and 4,981 to the other 18 external papers = 91% internal.

To address Q2, we are interested only in papers in the target research set to see how many of the cites to them (TCs) are by papers also within the target domain. For this study, we analyze just the highly cited (high CR) subset that are also identified as being in the research domain search set (e.g., 96 for NEDD, after re-searching WoS for TCs; recall our cleaning combined some CRs that should have been separate). This yields:

- NEDD: 96 papers cited 64,519 times in WoS, of which 34,982 are cites by our NEDD papers set (**54% internal)**
- HEV: 114 papers cited 8,522 times in WoS, of which 5,062 are cites by our HEV papers set (**59% internal)**

- DSSCs: 172 papers cited 73,752 times in WoS, of which 50,470 are cites by our DSSC papers set (**68% internal).**

A sidenote – these "Q2" analyses focus on the highly cited CRs in our target paper sets. In the search process, we gathered data on other of those highly cited CRs that are NOT in our target paper sets. For HEVs, for example, we note that some of those "external" CRs are very heavily cited outside the HEV paper set. Namely, of those 30 external CRs, 6 are cited over 1,000 times in WoS, in contrast to 20-51 cites by the HEV paper set. Five of those 6 were published in *Nature* or *Nature Materials*. For the DSSCs, one external paper stands way out with 38,255 cites in WoS vs. 288 by our DSSC paper set.

Discussion

The extent to which research domains stand apart – as proverbial "silos" – is important in terms of structures and processes to facilitate R&D productivity, creativity (i.e., through interdisciplinary exchange), and innovation [transferring knowledge toward technological development and applications (consider "translational" research to contribute to clinical practice in the biomedical arena)]. This study offers some new measures to offer new perspectives on such knowledge interchange, in addition to presenting some established tallies and maps.

The measures presented offer multiple perspectives. It is useful to see the subject categories (WOSCs) and Macro-Disciplines represented in a research domain (Figure 1). Comparing citation to publication behaviour helps see how these align for research domains (Table 2). We note that the journal is a more precise unit of analysis when compared with the journal grouping using WOS Categories (Rafols *et al*., 2010; Leydesdorff & Rafols, 2012; Leydesdorff, Carley & Rafols, in press). So, overlay maps based on journals promise more precise location of research activity.

But, such aggregations do not let us know to what extent a research endeavor is interdisciplinary in drawing together previously disparate knowledge. That is – is the research actually "integrative" (National Academies Committee, 2005) – in the sense of combining formerly separate knowledge (Rafols and Meyer, 2010)? That issue prompts exploration of the measures pursued here to distinguish amalgams of relatively separate research from novel interchanges of knowledge.

The current measures devised address the two research questions reasonably effectively. Regarding Q1 -- to what extent does the research domain reference external research? -- results indicate an ordering for our three sample sets, from DSSCs most internally oriented, to HEV research, to NEDD research, which appears most avaricious in drawing upon external research. Regarding Q2 -- to what extent do other fields cite the work of the research domain relative to its

self-citing? -- we find a similar progression, with DSSC (highly cited) papers receiving the most cites from within the field, to HEV research, to NEDD, receiving the highest proportion of cites by external papers. The ordering for the three sample cases fits our field knowledge gained in studying these emerging technologies. On the other hand, we recognize that changes in the search algorithms would alter the nature of the areas and the resulting scores – e.g., much NEDD research targets cancer and our search deliberately does not search on cancer terms per se; were such publications to be included, the self-containedness would surely change. Also, we focused on ~200 highly cited records in these analyses; again, a more or less inclusive set would alter the measures. Further research on sensitivity to such attributes would add considerable value to the measures.

These simple citation counting measures offer promise, but are no means sufficiently tested or refined for general adoption. We note that tabulation is not automatic because CRs are not reported completely consistently. The clumping algorithm we used worked pretty well at consolidating CR variations, but warrants further exploration. We believe clumping errors, on balance, tend to overestimate the within-domain counts. "Giants" (papers with huge numbers of cites received) pose an additional challenge in deciding how best to treat them in arriving at measures of central tendency. We report means (averages) here as the target sets were not unduly sensitive to such papers, but would look for a more robust strategy (perhaps focusing on something like the 10-90th percentile range), but alternatives remain to be compared.

The measures introduced here could certainly be "translated" to other scales. Of special interest might be to explore how these score for research threads – much smaller, more discrete research area characterizations (Boyack et al., 2012). And conversely, Boyack et al. (2009) have mapped citation "flows" into and out of larger aggregations (namely, chemistry and related subject categories, in that paper) and how those have changed over time. In between, examination of the knowledge flows (indicated by citations in and out) among sub-fields within a research domain could prove informative in better managing R&D. For instance, might a hypothetical case be made that DSSC researchers could benefit by being more attentive to dye research from outside their field? Going further, one reviewer noted that it would be interesting to examine how self-containedness relates to expansibility and sustainability of research domains. One could imagine studying such phenomena at various scales, from research threads on up.

Many factors could affect self-containedness. Emerging research areas would seem apt to transition from "nothingness" to a degree of awareness of others' research as relevant, and, perhaps, beyond -- to phases in which specialization within the area grows and sub-fields branch off. Thinking about DSSCs, HEVs, and NEDD – differences in research norms as one spans different domains would

seem a factor. These three emerging technologies also seem inherently different in the degree to which they would focus externally or not – NEDD addresses manifold applications of multiple technologies, whereas DSSCs all address one family of related means to achieve a singular end of effective solar energy conversion. HEVs are embedded in innovation processes engaging complex transportation systems and policies that seem wider in scope than DSSC research issues at this time. We offer our simple self-containedness measures as tools to study such factors.

Multiple perspectives are highly desirable in characterizing the "self-containedness" of a research domain. We offer the present simple calculations as easy-to-understand indicators. We see nice potential in multiple mappings, noting the journal-based overlay possibilities to locate research domains in science, and also in terms of core journals and their relatedness (Leydesdorff and Rafols, 2012; Leydesdorff et al., to appear). We see promise in mapping the respective generations. For instance, take NEDD. We intend to generate co-citation maps to help characterize the "knowledge source" domain – separately for the internal and external CR sets. Given that we have gathered their WoS records, we can try these based on first or all authors, and/or individual papers, and/or on institutions. Conversely, we also want to examine the co-citing patterns, but those pose greater challenge in data gathering.

Another extension would be to compare these self-containedness metrics with other indicators for given research areas. For instance, one could compare these three areas in terms of co-author or co-citation network densities. Bibliographic coupling comparisons would also be quite interesting. Furthermore, it could be fruitful to pursue bibliographic coupling maps and measures to see if those distinguish sub-systems within a given research domain (e.g., who cites different segments of the CR space?).

We offer the present simple measures and early results to stimulate consideration of such possible indicators of research domain "porosity." We see these as dialogue and experimentation starters, not finished products.

Acknowledgements

This research was undertaken at Georgia Tech drawing on support from the National Science Foundation (NSF) Science of Science Policy Program -- "Revealing Innovation Pathways" (Award No. 1064146). The findings and observations contained in this paper are those of the authors and do not necessarily reflect the views of the National Science Foundation.

References

Arora, S.K., Porter, A.L., Youtie, J. & Shapira, P. (2013). Capturing new developments in an emerging technology: An updated search strategy for identifying nanotechnology research outputs, *Scientometrics*, 95 (1), 351-270.

Boyack, K. W., Borner, K. & Klavans, R. (2009). Mapping the structure and evolution of chemistry research. *Scientometrics*, **79**, 45-60.

Boyack, K.W., Klavans,R., Small, H.,& Ungar, L. (2012). Characterizing emergence using a detailed micro-model of science: Investigating two hot topics in nanotechnology, *Portland International Conference on Management and Engineering Technology (PICMET)*, Vancouver.

Carley, S. & Porter, A.L.(2012). A forward diversity index, *Scientometrics*, 90 (2), 407–27.

Leydesdorff, L., Carley, S. & Rafols, I. (2013). Global maps of science based on the new Web-of-Science categories, *Scientometrics,* 94 (2), 589-593.

Leydesdorff, L. & Rafols, I. (2009). A Global map of science based on the ISI Subject Categories. *Journal of the American Society for Information Science and Technology*, 60 (2), 348-362.

Leydesdorff, L. & Rafols, I. (2012). Interactive overlays: A new method for generating global journal maps from Web-of-Science data, *Journal of Informetrics,* 6 (2), 318-332.

Leydesdorff, L., Rafols, I. & Chen, C. (to appear). Interactive overlays of journals and the measurement of interdisciplinarity on the basis of aggregated journal-journal citations. *Journal of the American Society for Information Science and Technology.*

Ma, T., Porter, A.L., Ready, J. Xu, C., Gao, L., Wang, W. & Guo, Y. (under revision). A technology opportunities analysis model: applied to Dye-Sensitized Solar Cells for China, *Technology Analysis and Strategic Management*.

National Academies Committee on Facilitating Interdisciplinary Research, Committee on Science, Engineering and Public Policy (COSEPUP) (2005). *Facilitating interdisciplinary research*. (National Academies Press, Washington, DC).

Porter, A.L., Cunningham, S.W. & Sanz, A. (to appear). Extending the FIP (Forecasting Innovation Pathways) approach through an automotive case analysis, *Portland International Conference on Management and Engineering Technology (PICMET)*, San Jose, California, 2013.

Porter, A.L. & Rafols, I. (2009). Is science becoming more Interdisciplinary? Measuring and mapping six research fields over time, *Scientometrics*, 81(3), 719-745.

Porter, A.L., Roessner, J.D. & Heberger, A.E. (2008), How interdisciplinary is a given body of research?, *Research Evaluation*, 17 (4), 273-282.

Rafols, I. & Meyer, M. (2010). Diversity and network coherence as indicators of interdisciplinarity: case studies in bionanoscience. *Scientometrics*, 82(2), 263-287.

Rafols, I., Porter, A.L. & Leydesdorff, L. (2010). Science overlay maps: A new tool for research policy and library management. *Journal of the American Society for Information Science and Technology*, 61 (9), 1871–1887.

Stirling, A. (2007). A general framework for analysing diversity in science, technology and society. *Journal of The Royal Society Interface*, 4 (15), 707-719.

Wagner, C.S., Roessner, J.D., Bobb, K., Klein, J.T., Boyack, K.W., Keyton, J., Rafols, I. & Borner, K. (2011),. Approaches to understanding and measuring interdisciplinary scientific research (IDR): A review of the literature, *Journal of Informetrics*, 5165, 14-26.

Zhou, X., Porter, A.L., Robinson, D.K.R. & Guo, Y. (to appear), Analyzing Research Publication Patterns to Gauge Future Innovation Pathways for Nano-Enabled Drug Delivery, *Portland International Conference on Management and Engineering Technology (PICMET)*, San Jose, California, 2013.

Zhou, X., Porter, A.L., Robinson, D.K.R. & Guo, Y. (to appear), Patent and Publication Comparison for One Emerging Industries -- Nano-Enabled Drug Delivery, *14th International Society of Scientometrics and Informetrics (ISSI) Conference*, Vienna, 2013.

Zitt, M. & Small, H. (2008). Modifying the journal impact factor by fractional citation weighting: The audience factor, *Journal of the American Society for Information Science and Technology*, 59 (11), 1856-1860.

Appendix

The DSSC search has been developed by Ying Guo and Tingting Ma through a series of analyses over the past few years (Ma et al., under review). It was rerun for the present study on 22 Jan., 2013, for 1991-2012 in WoS (including SCI-expanded, SSCI, CPCI-S & SPCI-SSH) The main search phrase is:

- TS= (((dye-sensiti*) or (dye* same sensiti*) or (pigment-sensiti*) or (pigment same sensiti*) or (dye* adj sense)) same ((solar or Photovoltaic or photoelectr* or (photo-electr*)) same (cell or cells or batter* or pool*)))

Additional phrases used:

- TS= (((dye- Photosensiti*) or (dye same Photosensiti*) or (pigment-Photosensiti*) or (pigment same Photosensiti*)) same ((solar or Photovoltaic or photoelectr* or (photo-electr*)) same (cell or cells or batter* or pool*)))
- TS= (((dye- optoelectri*) or (dye same optoelectri*) or (pigment- optoelectri*) or (pigment same optoelectri*) or (dye- opto-electri*) or (dye same opto-electri*) or (pigment- opto-electri*) or (pigment same opto-electri*)) same ((solar or Photovoltaic or photoelectr* or (photo-electr*)) same (cell or cells or batter* or pool*)))
- TS = (((dye and (conduct* or semiconduct*)) same electrode*) and electrolyte*)

This yielded **8919 records**.

Search for HEVs was conducted as part of an exercise in "FIP" -- Forecasting Innovation Pathways (Porter et al., to appear). The search used here was run on 22 Jan., 2013 for 2000-2012 in WoS (including SCI-expanded, SSCI, CPCI-S & SPCI-SSH) for:

ts= ((electric or hybrid) near/2 (vehicle or vehicles or automobile or automobiles or car or cars))

It yielded 7720 records, from which articles (3496) and proceedings papers (4384), for a total of **7323 records** were downloaded. The HEV search used in the FIP exercise was iteratively devised and incorporated many subsystem search modules (e.g., on mechanical energy recovery, thermal management, batteries, electric motors for EVs, fuel cells for cars, electromagnetic brakes, flywheels, hydrogen storage, lightweight materials & vehicles, etc.

The NEDD search in WoS was based on an earlier search strategy from 2009, iteratively revised with expert review over about 9 months in 2012 (Zhou et al., to appear). The basic strategy is to combine *Delivery* terms AND *Nanotechnology* terms AND (*Pharma* terms or highly selective *Target* terms). In the 2009 search, a major component concerned the Target of cancer; this was not explicitly included here. This search included 7 modules. The following were addressed to a Nano subset of WoS (Arora et al., 2013) OR to records containing alternative terms – (viral* OR colloid* OR dendrimer*):

- TS=((deliver* or vehicle* or carrier* or vector* or "control* releas*") Near/4 (Drug* or pharmacy)
- TS=((deliver* or vehicle* or carrier* or vector* or "control* releas*" or transduct* or transfect* or transport* or translocat*) Near/4 agent*)
- TS=((deliver* or vehicle* or carrier* or vector* or "control* releas*" or transfect*) Near/4 formulation*)
- TS= (deliver* or vehicle* or carrier* or vector* or treat* or therap* or "control* releas*" or transduct* or transfect* or transport* or translocat*) Near/4 (DNA or gene)

The following were addressed to WoS:

- TS=((deliver* or vehicle* or carrier* or vector* or treat* or therap* or "control* releas*" or transduct* or transfect* or transport* or translocat*) Near/4 (siRNA or "short interfering RNA"))
- TS= (deliver* or vehicle* or carrier* or vector* or treat* or therap* or "control* releas*" or transduct* or transfect* or transport* or translocat*) Near/4 (Dox or Doxorubicin*)
- TS=((deliver* or vehicle* or carrier* or vector* or treat* or therap* or "control* releas*"or transfect*) Near/4 ("RNA interference" or RNAi))

This yielded **61,465 records.**

A METHOD FOR TEXT NETWORK ANALYSIS: TESTING, DEVELOPMENT AND APPLICATION TO THE INVESTIGATION OF PATENT PORTFOLIOS (RIP)

Luciano Kay[1]

[1] *luciano@cns.ucsb.edu*
Center for Nanotechnology in Society (CNS), University of California Santa Barbara, Santa Barbara, CA (USA)
Georgia Tech Program in Science, Technology and Innovation Policy (STIP), Georgia Institute of Technology, Atlanta, GA (USA)

Abstract
This research explores Text Network Analysis (TNA) as an alternative method to analyze scientific and patent literature. This paper introduces the TNA method and presents some results of algorithm testing, development and implementation. We exemplify the practical implementation of the method with the analysis of a set of patent records by Japanese companies in the field of nanotechnology applied to energy storage solutions between 2000 and 2009. Although this is a research in progress paper, we are able to identify some features and potential issues in the use of the TNA method to investigate science, technology and innovation topics using scientific publication and patent data. Improvements and calibration of this method are under development.

Conference Topic
Visualisation and Science Mapping: Tools, Methods and Applications (Topic 8) and Technology and Innovation Including Patent Analysis (Topic 5).

Introduction
A number of approaches have been developed to analyze scientific publication and patent document data and investigate a broad range of phenomena related to science, technology and innovation topics. Those approaches involve, for example, metrics to identify most frequent or relevant terms found in the academic and patent literature (e.g. word frequencies and term frequency–inverse document frequency TF-IDF) and term clumping and identification of concepts using "topic modelling" and clustering techniques such as Principal Component Analysis (Porter et al., 2012).
This research explores Text Network Analysis (TNA) as an alternative method to analyze scientific and patent literature. TNA is a set of methods to extract meaning and identify pathways for meaning circulation from text corpora based upon conceptual linkages (Paranyushkin, 2011). The TNA analysis draws on the analysis of co-occurrence of words-phrases in the text and the application of

social network analysis techniques. It has been applied, for example, to the analysis of transcripts (Broniatowski, 2012) or rhetoric in scientific publications (Long, 2012) (this is generally not academic literature.) Among the alleged advantages of this approach there are the ability to effectively identify the most influential concepts that produce meaning and the possibility of performing comparative analysis of different kinds of texts.

A method for Text Network Analysis

In this paper we present some results of the testing, development and implementation of TNA analysis applied to patent literature. This process follows four main steps which are further developed in the rest of the paper:

Data extraction. This step involves the identification of "terms" or sets of words-phrases from the source text. Herein, we start with a set of patent application records, we merge both title and abstract fields and extract "NLP phrases" using the Natural Language Processing (NLP) routine available in VantagePoint text-mining software. Then we extract word-phrases (hereafter, simply "terms") using another automation macro developed by the author that further cleans up the list of NLP phrases and discards irrelevant terms.

Text network creation. TNA requires representing the source text in an adjacency matrix format to identify concepts and develop metrics. We do this using the co-occurrence matrix creation routine available in VantagePoint. The resulting undirected, weighted matrix is the adjacency matrix used as an input for the next steps.

Clustering. The terms are clustered to form more or less homogeneous groups, i.e. clusters of terms that are better connected among them than with the rest of the co-occurrence network formed by the terms. We apply the Markov Cluster (MCL) algorithm to cluster terms according to their co-occurrence in titles and abstracts of patent documents. The MCL algorithm is a scalable unsupervised cluster algorithm for networks based on simulation of stochastic flow in graphs (more information on MCL is at http://micans.org/mcl/).

Concept analysis. We apply a measure of betweenness centrality to the clusters found in previous steps to identify key concepts or themes within each set.

Algorithm testing and development

To test and develop this method we use data from EPO Patstat on patent applications in the nanotechnology for energy storage field. Energy storage technologies are those defined by the OCDE patent search strategy in terms of IPC classes and ECLA codes. We use the keyword-based definition of nanotechnology described in Porter et al. (2008). We use a subset of patents from all patent authorities filed by Japanese companies between 2000 and 2009. This dataset comprises 262 patent applications (only patent documents in English language are included.) The VantagePoint NLP phrases creation routine applied to titles and abstracts and further automated work to clean up terms yield 304 words/phrases (Table 1).

Table 1. Top-15 terms in the Japanese corporate nanotechnology for energy storage patent applications dataset (2000-2009)

	Terms	*Patent records*
1	Carbon Nanotubes	45
2	Dye-Sensitized Solar Cell	40
3	Lithium Ion Battery	34
4	Electrolyte	27
5	Fuel Cell	21
6	Fullerene	21
7	Non-Aqueous Electrolyte	21
8	Carbon Nanofibers	20
9	Carbon Fibers	17
10	Secondary Battery	15
11	Electrodes	14
12	Counter Electrode	9
13	Electrochemical Capacitors	9
14	Electrode Substrate	8
15	Hexagonal Carbon	8

Using the co-occurrence matrix creation routine available in VantagePoint, we create an undirected, weighted adjacency matrix (i.e. cell values contain the number of records in which two given terms appear together.) Visual inspection of this matrix already suggests the concentration of linkages (i.e. co-occurrences) between the most frequent terms rather than their more homogeneous distribution across the whole set of terms, which is likely the result of a set of terms focused on a very specific topic.

We cluster terms using the MCL algorithm and test different parameters using a sub-sample of the dataset with only the top-50 words/phrases in terms of patent records (to speed up processing.) Three main parameters are set: expansion, inflation, and iterations (Macropol, 2009). We found that the MCL process converges before 10 iterations (sooner than suggested by the algorithm author) and the variation of this parameter above this level does not affect the clustering results. Expansion and inflation do influence more importantly (and in opposite directions) clustering outputs (Table 2a and Table 2b).[113] We find that setting expansion=2 and inflation=2—their minimum value—lead to more satisfactory results (i.e. not too much granularity and overlap between clusters not significant.) We verify that the use of weighted matrices (rather than Boolean matrices) leads to less overlap between clusters. While the original implementation of the MCL algorithm (Macropol, 2009) suggests adding self loops in adjacency matrices to avoid the strong effect of odd powers of expansion, we purposely exclude self loops. The result of this is slightly more homogeneous clusters with less concentration of terms in the biggest cluster.

[113] This sensitivity analysis is performed using the top-50 terms according to the number of patent records they relate to. Total vertices=number of terms; when clustered vertices is larger than total vertices, cluster overlap.

Table 2a. Sensitivity of MCL algorithm to expansion parameter

Expansion	*Inflation*	*Total vertices*	*Total clusters*	*Total clustered vertices*	*Median cluster size*
2	2	50	12	85	4.5
3	2	50	3	50	4
4	2	50	3	50	4
5	2	50	3	50	4

Table 2b. Sensitivity of MCL algorithm to inflation parameter

Expansion	*Inflation*	*Total vertices*	*Total clusters*	*Total clustered vertices*	*Median cluster size*
2	2	50	12	85	4.5
2	3	50	15	51	3
2	4	50	19	50	2
2	5	50	22	50	1

We then compute the normalized betweenness centrality (Brandes, 2001) for terms within each cluster to identify to most relevant concepts. This measure indicates how often a node appears between any two random nodes in the network. The higher the betweenness score of a term, the more influential it is because it functions as a junction for communication within the network (Paranyushkin, 2011). We find however that this measure tends to be zero or near zero for most of the terms within each cluster as they are only linked with a small, well-connected set of terms within the cluster (we think this is likely a dataset-specific phenomenon.) The measure still helps to identify the most influential terms within each cluster.

Implementation results

We applied the TNA to titles and abstracts of a set of 262 patent applications by Japanese companies between 2000 and 2009. The analysis yields 32 clusters or groups of terms (6 of them contain only one term.) A cursory inspection of the terms in each cluster indicates that some groups are somewhat heterogeneous, which may be evidence of more complex relationships between different technology concepts that cannot be appreciated by an inexpert observer. We also identify words/phrases that need further clean up—the data creation pre-processing step is automated—and may affect the clustering process to some extent.

We observe that terms tend to be clustered into few main groups (or often, one main group) which is more likely the result of the focus of the original dataset (i.e. the dataset covers a specific topic, nanotechnology applied to energy storage.) Using a dataset with a specific focus is still interesting from the point of

view of development and testing as the author is able to better interpret results related to data and topic he is familiar with.
The top-10 clusters represent about 80% of the co-occurrence linkages and almost the same proportion (79%) of patent records in the dataset (Table 3). In this case, the most densely connected clusters are those linked to technologies such as Solar cells and Battery components. These account for about 40% of linkages between terms. Also the proportion of patent records in clusters such as Solar cells and Lithium Ion Batteries suggests more activity in those areas.

Table 3. Top-10 clusters of related concepts in the application of TNA analysis to a patent portfolio of Japanese companies

Cluster[a]	*# terms*	*% total edges*	*% total records*[b]	*Top-3 terms (norm. betweenness in parenthesis)*
Solar cells	62	19.09	26.52	Dye-Sensitized Solar Cell (0.49), Electrolyte (0.37), Electrochemical Capacitors (0.06)
Battery components	37	17.71	13.66	Carbon Nanofibers (0.40), Non-Aqueous Electrolyte (0.38), Electrodes (0.04)
Nano-materials	40	9.94	18.08	Carbon Nanotubes (0.93), Long-Chain Carbon Nanotubes (0.01), Electrical Conductivity (0.00)
Fuel cells	32	9.83	12.05	Fuel Cell (0.83), Electrochemical Capacitors (0.13), Polymer Electrolyte (0.02)
Lithium Ion Batteries	34	9.37	20.09	Lithium Ion Battery (0.65), Nanoparticles (0.19), Carbon Atoms (0.06)
Hydrogen cells	25	8.11	8.84	Fullerene (0.82), Hydrogen Atoms (0.00), Hydrogen Molecules (0.00)
Carbon composites	9	2.06	7.23	Carbon Fibers (0.46), Carbon Composite (0.25), Carbon Particles (0.00)
Nickel alloys	6	1.71	1.21	Electrochemical Electrode (0.00), Nanostructures (0.00), Nickel Nanocrystal (0.00)
Equipment	6	1.71	0.8	Electrolyzers (0.00), Flowmeter (0.00), Hydrogen Generator (0.00)
Secondary batteries	9	1.49	7.63	Secondary Battery (0.75), Hexagonal Carbon (0.18), Truncated Conical Tubular Graphene (0.18)

Notes: a. cluster name assigned by the researcher. b.percentage of total records exceed 100% due to terms that appear in more than one records and clusters that overlap. Table shows only the top-10 clusters in terms of share of total network edges.

The calculation of the normalized betweenness score helps to identify those terms that are the most central in each group. Their interconnectedness with several other peripheral terms within each cluster suggests a role for them in the conceptual interpretation of the groups of terms. Cluster names can be created based on these most central terms. We find, however, that most of the terms have very low betweenness centrality scores, indicating the existence of weakly connected groups that do not have a distinctive central term.

Network graph visualization is also helpful in this regard. We use the software Gephi to visualize the network of all clusters discovered in this dataset (Figure 1). The graph shows closer connections between technologies related with Solar Cells, Lithium Ion Batteries, and Nanomaterials (particularly, "carbon nanotubes"). This focus of the patenting activity of Japanese companies in this field coincides in general terms with the results of the analysis of leading countries as investigated with other methods (Kay & Appelbaum, 2012).

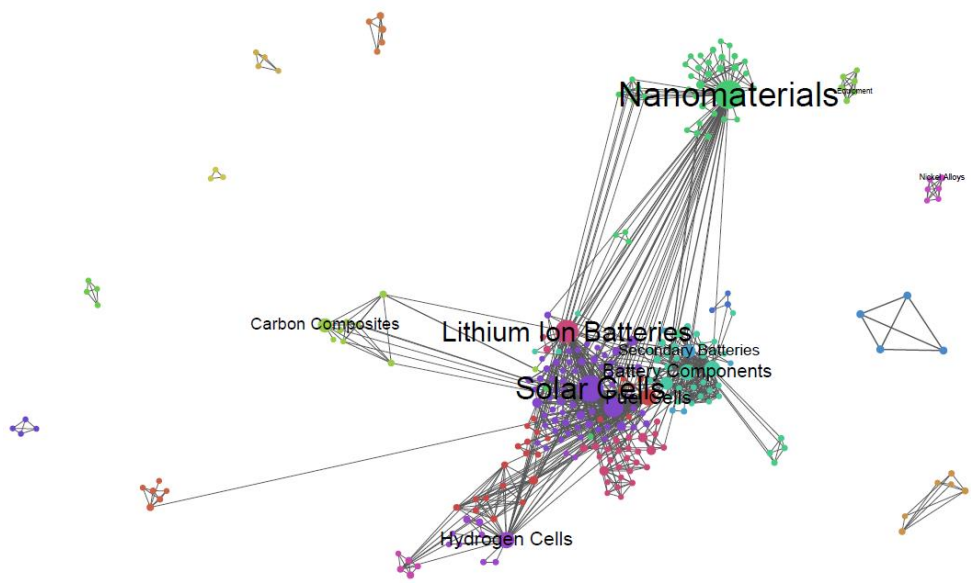


Figure 1. Network graph visualization of terms and clusters discovered using TNA approach in nanotechnology energy storage patent applications by Japanese companies (2000-2009)[114]

[114] The graph shows labels for top-10 clusters only. The size of labels represents the number of patent records associated with the cluster. Node colour represents the cluster each term is associated with. We use Gephi and OpenOrd layout algorithm.

Conclusions and next steps

This paper introduces a method for Text Network Analysis and presents some results of algorithm testing, development and implementation. Preliminary findings show some interesting features of this method. Most importantly, it offers a number of parameters that can be calibrated to improve the method and make it more effective to investigate scientific and patent literature. It also allows the creation of both quantitative indicators and visualizations that may help to interpret results and discover themes and conceptual relationships. Regarding the specific application of the method to patent portfolio analysis, TNA may complement other analysis that draw on more conventional categories such as IPC classes. We find that the interpretation of TNA results, however, is not straightforward, particularly when there are an increasing number of terms and overlapping clusters. In this regard, the existence of groups of terms that are weakly connected or a number of single-element clusters suggests that some pruning may be beneficial before applying this routine to focus only on the most important terms.

Improvements and calibration of this method are under development. Although there are only a few parameters to calibrate the MCL algorithm, its implementation demands significant hardware resources when analyzing big datasets. This results in increased testing and calibration time. Next steps should include algorithm improvement to increase speed of analysis; application to datasets that do not focus on specific topics (an example would be all patent applications by Japanese companies during a certain time period;) direct application to patent document text instead of application to either titles and abstracts or their NLP phrases; and, comparison with other methods for topic discovery such as Principal Component Analysis.

Acknowledgments

This work has been supported by the National Science Foundation (Grant No. SES 0938099 and SES 0531184). Any opinions, findings, conclusions or recommendations expressed are ours and do not necessarily reflect those of the National Science Foundation. We conducted our research under the auspices of the UCSB's Center for Nanotechnology in Society (www.cns.ucsb.edu).

References

Brandes, A. (2001). Faster Algorithm for Betweenness Centrality. *Journal of Mathematical Sociology* 25(2):163-177.

Broniatowski, Magee (2012). Studying Group Behaviors. A tutorial on text and network analysis methods. *IEEE Signal Processing Magazine*, 22-32.

Kay, L. & Appelbaum, R. (2012). How do companies embrace emerging technologies? The case of nanotechnology and energy storage applications in China. 2012 Conference on Patent Statistics for Decision Makers (PSDM). Paris, France, November 28-29.

Long, Seth (2012). Concept clusters in Rhetoric Society Quarterly. Retrieved January 25, 2013 from: http://technaverbascripta.wordpress.com/2012/10/08/text-network-analysis-1-concept-clusters-in-rhetoric-society-quarterly/

Macropol, Kathy (2009). Clustering on Graphs: The Markov Cluster Algorithm (MCL)

Paranyushkin, Dmitry (2011). Identifying the Pathways for Meaning Circulation using Text Network Analysis. Retrieved January 25, 2013 from: http://noduslabs.com/publications/Pathways-Meaning-Text-Network-Analysis.pdf

Porter, Alan L.; Newman, David; Newman, Nils C. (2012). Text Mining to Identify Topical Emergence: Case Study on Management of Technology.

Porter, A. L., Youtie, J., Shapira, P., & Schoeneck, D. J. (2008). Refining search terms for nanotechnology. *Journal of Nanoparticle Research*, 10, 715-728.

MISFITS? RESEARCH CLASSIFICATION IN RESEARCH EVALUATION: VISUALIZING JOURNAL CONTENT WITHIN FIELDS OF RESEARCH CODES

Gaby Haddow[1] and Ed Noyons[2]

[1] *g.haddow@curtin.edu.au*
Curtin University, Dept of Information Studies, GPO Box U1987, 6845 Perth (Australia)

[2] *noyons@cwts.leidenuniv.nl*
Leiden University, Centre for Science and Technology Studies (CWTS), PO Box 905, 2300 AX Leiden (The Netherlands)

Abstract
The Australian research evaluation model uses a classification scheme to assign Fields of Research (FoRs) to individual researchers and journals, and to define assessment panels. Eligible journals for assessment are listed and assigned between one and three FoR codes. A high proportion of journals in the list of over 22,000 titles are assigned a single FoR code only. This paper explores the implications of classifying research outputs using the FoR code mechanisms. Eight datasets of title and abstract data from journals assigned a single FoR were mapped using VOSviewer. Four of the datasets were in science fields and the other four were humanities and social sciences fields. The maps and extracted terms for each journal set were examined for overlap with other FoRs. Sizeable overlaps with other FoR codes were observed in the content of three of the sciences fields' datasets. Weaker overlaps with other FoR codes were found in the humanities and social sciences datasets. The findings suggest that the assignment of FoR codes to journals has the potential to disadvantage researchers and their organisational units.

Conference Topic
Science Policy and Research Evaluation: Quantitative and Qualitative Approaches (Topic 3) and Visualisation and Science Mapping: Tools, Methods and Applications (Topic 8).

Introduction
Research evaluation serves a variety of purposes. At the macro level of countries and institutions, it can encourage higher quality and quantity of research output; identify areas of strength, weakness and duplication; and provide an accountable and transparent method of supporting research activity (Box, 2010; Butler, 2010; European Science Foundation, 2012). At the micro level, research evaluation provides an opportunity for organisational units and individuals to prove their value and contribution to their institution. Ideally, the mechanisms applied in research evaluation deliver outcomes that enable macro level decision making while recognising the efforts of those at the micro level.

An important factor in achieving a match between macro and micro level needs is the method used to classify research by field. Most national research evaluation frameworks appoint panels of experts that assess research outputs submitted by participating institutions. In the United Kingdom (Higher Education Funding Council for England, 2005) and New Zealand (Tertiary Education Commission, 2012) these submissions are based on selecting outputs that align with defined parameters of expert disciplinary panels. If a panel deems a submission out of scope or a submitting institution requests it, then referral to another assessor may occur. This mechanism allows for cross-disciplinary research to be assessed by the most appropriate panel or expert, while retaining a connection between an individual or group submission and their research. The Australian research evaluation framework, Excellence in Research for Australia (ERA), also uses panels. Unlike the models described above, these panels are defined by a classification scheme, the Australian and New Zealand Standard Research Classification (ANZSRC) (Australian Bureau of Statistics, 2008). In addition, classification is extended to the assignment of Fields of Research (FoR) codes to researchers and research outputs (Australian Research Council, 2011). Over 22,000 journals (the ERA 2012 Journal List), deemed eligible outputs, are assigned between one and three FoR codes. Some allowance for multidisciplinarity is built into the ERA with an MD code and broader two-digit codes, however journals assigned four digit FoR codes comprise the vast majority (85%) and over 50% of all journals in the ERA list are assigned a single code. By using FoR codes to classify researchers, research outputs and assessment panels, the Australian research evaluation model has created a system in which a classification scheme is the overarching structure. In effect, it has the potential to separate researchers from their research output.

Background

Classification schemes, such as the Universal Decimal Classification and the extensive Medical Subject Headings (MeSH), are well known and well researched. Schemes applied specifically to research activities are more recent and, as a result, research into their impact in research evaluation is limited. The growth of research evaluation activities internationally (ESF, 2012; Goldfinch & Yamamoto, 2012) is raising the profile research classification, however. A European Science Foundation (ESF) forum discussed several schemes to identify key issues which might improve the capacity to compare research between organisations and countries. An ESF working document (2011) stressed the importance of consistent coding and mapping between different classification schemes to provide "a powerful tool, particularly when applied to national portfolios to plan to address gaps and new opportunities" (p. 17). The potential for text mining using title and abstract data was also discussed as a means to classify research. Drawing on the working document, a later ESF report notes "one of the greatest challenges in research evaluation is to connect information to a researcher, a grant, an output" (2012, p. 10).

Accurate classification is vital to achieve a primary objective of evaluation activities: "to provide tools in decision making processes about the allocation of research funds" (Moed, 2005). In many national systems, research is informally classified by the researcher through selection of an assessment panel to submit their work for peer review, based on a judgment that the subject content of their research corresponds with a panel's expertise. This is the process for research evaluation in the United Kingdom and New Zealand, and both limit the number of research outputs that can be submitted by a researcher (Goldfinch & Yamamoto, 2012). In contrast, submissions to the ERA include all eligible research outputs over a specified period, individual researchers must select between one and three FoR codes to represent their research activities, and journal article outputs are bound by a pre-determined classification. That is, articles are assigned the same FoR code or codes as the journal in which they are published, unless a researcher can argue that over 66% of the content relates to a different FoR. Discussing the FoR code mechanism for ERA journals, it has been noted: "accuracy at the article level will only be achieved if a) the FoRs are allocated to journals accurately and consistently, and b) individual articles conform to the subject focus of a journal as it is expressed in the FoRs" (Bennett, Genoni & Haddow, 2011, p. 89). These researchers found 46.8% of codes assigned by authors to their articles did not align with the journals' codes as assigned in the ERA.

Basically there are three types of problems with a journal classification: resolution of the scheme, inter-disciplinarity and scope of journals. The first type refers to the struggle to describe a proper structure of science. Clearly, 'natural sciences' is too broad to be a meaningful class, but how about 'physics'. In many cases that is still too broad, but what is an appropriate level. There is no real reference to the 'real world', no definite scheme. The second type is the one we address in this paper. Journals may act at the interface of one 'field' to the other but should the output and impact of such journals be divided equally over the two fields at stake? As long as we use journals as the entity to classify we will need to answer this. Finally, the third type relates to the problem of multi-disciplinary journals but also to journals that have a broader scope than specialized journals. Journals like 'The Lancet' are general medicine journals and as such matching a scheme where 'medicine' is the highest resolution reached. But when a higher resolution is needed, such journals are problematic.

Classification schemes, such as the ANZSRC, are designed to create clearly defined categories, enabling consistency in their application. For the ERA model of research evaluation to operate accurately and equitably, a researcher and their research outputs need to be classified under the same FoR codes. A difference between a researcher's selected FoR codes and the codes assigned to their journal article outputs may result in the researcher's work being assessed under codes that represent a different field, and possibly a different organisational unit (Kwok, 2013).

This preliminary study explored the implications of classifying journal articles using the FoR code mechanisms of the ERA. It mapped the content of journals assigned a single FoR code and analysed the primary clusters and terms evident in the content. By examining this content against FoR code definitions and exclusion criteria, it is possible to assess the correspondence between assigned codes and journal content. The results are then considered in relation to the implications for researchers and organisational units.

Methods

The data required for the study were article titles and abstracts from journals assigned a single four-digit FoR code in the ERA 2012 Journal List (Australian Research Council, 2012). In order to gather sufficient content for analysis, the study identifed large journals sets with the highest proportion of titles assigned one four-digit FoR code only. Codes with less than 200 journals assigned the code as their first FoR were excluded to ensure large journal sets formed the base data for the next stage. The large journal sets were then examined to identify those with 60% or more titles assigned the single code and an ISSN search of the CWTS Web of Science database was performed to ensure title and abstract content for the journals was available. Journal sets with coverage by the CWTS database of around 40% or more for science fields and over 20% for humanities and social science (HSS) fields comprised the final sample. This resulted in eight FoR code journal sets for analysis, equally divided between sciences and HSS fields. Table 1 data presents these data for the selected journal sets.

Table 1. Selected journal sets for analysis

Field of Research	*FoR code*	*Journals assigned code (No.)*	*Journals assigned single code*		*CWTS coverage single code*	
			(No.)	*(%)*	*(No.)*	*(%)*
Pure Mathematics	0101	502	371	73.9	206	55.5
Zoology	0608	355	247	69.6	143	57.9
Nursing	1110	270	163	60.4	64	39.3
Pharmacology & Pharmaceutical Sciences	1115	418	288	68.9	152	52.8
Performing Arts & Creative Writing	1904	366	226	61.7	59	26.1
Literary Studies	2005	1070	661	61.8	166	25.1
Archaeology	2101	362	244	67.4	53	21.7
Historical Studies	2103	1077	656	60.9	180	27.4

Title and abstract data for articles in the eight journal sets, published between 2008 and 2010, were downloaded from the CWTS database. Due to variations in the number of journals in a set and the availability of abstract data, the datasets differed substantially in size. For example, Pure Mathematics and Historical Studies produced datasets with over 40,000 items, while Archaeology had around

5,000. Exceeding all others was the dataset for Pharmacology, with over 79,000 items. These data were analysed for content using the VOSviewer software (version 1.5.3). The VOSviewer is an open source application to analyze and visualize network data. The application is able to identify groups within network data using a modularity-based clustering technique. More details about the method and technique are in Waltman, Van Eck and Noyons (2010) and Van Eck & Waltman (2010). To allow for dataset size variations and aiming to produce maps with clearly defined clusters, different term occurrence thresholds were applied in the VOSviewer analyses. A thesaurus file for VOSviewer was created for each dataset to eliminate publishers' names and to ensure consistency for different spelling (eg. theatre/theater) and for synonyms (eg. sixteenth century/16th century). The software enabled visualization of the journal sets' content and provided a list of the main terms and noun phrases used by authors in their titles and abstracts. The most relevant terms are selected by an algorithm (Van Eck et al., 2010) taking the discriminative strength of terms into account. Together the selected terms do not cover the entire set of publications but always a representative part. The main clusters in the VOSviewer map and the list of extracted terms were examined against the ANZSRC definitions and exclusions for the FoR codes. Additional resources about a field's coverage and parameters, such as subject dictionaries, were consulted to aid the analysis.

Findings

Pure Mathematics 0101

The VOSviewer map for Pure Mathematics used a threshold of 25 term occurrences, from which the software generated 1527 relevant terms. The dataset was drawn from 206 journals and comprised around 43,000 items. Five clusters are evident in the map (Figure 1) and represent content relating to algebra, equations and components, number theory, functional analysis and mathematical analysis, and geometry. These clusters appear to conform to the definitions of the FoR code. However, a number of terms that occur in the map are closely related to mathematical physics, which is classified with a different FoR code and specifically excluded from the 0101 code. A particularly high occurrence of the noun phrase 'Hilbert space' (398) and occurrences of variants on the phrase were observed in the central cluster. The terms, 'Schrodinger operators' and 'Hamiltonian' (with variant noun phrases), were also found in the map with relatively high occurrences (163 and 188, respectively).

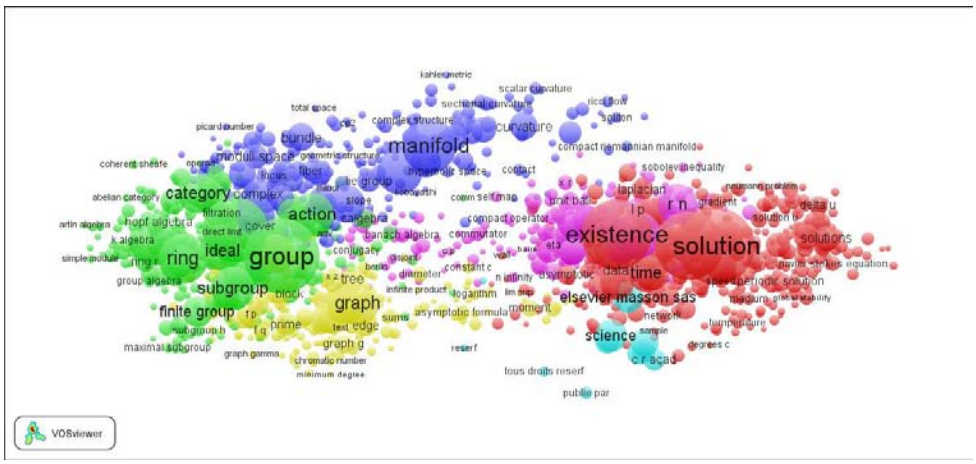


Figure 1. VOSviewer map of journal content for Pure Mathematics (FoR code 0101)

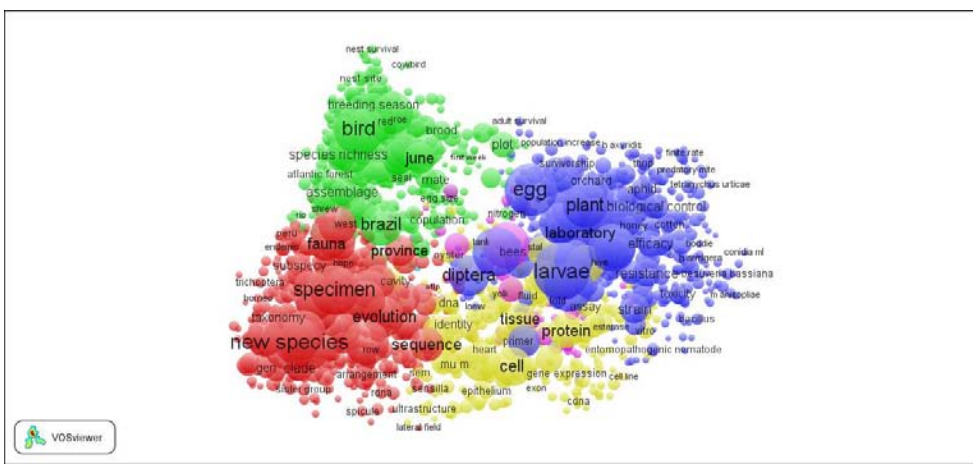


Figure 2. VOSviewer map of journal content for Zoology (FoR code 0608)

Zoology 0608

Using a threshold of 25 term occurrences, VOSviewer generated 2078 relevant terms to map in the Zoology journal set of over 28,000 items from 143 journals. The map (Figure 2) displays a tightly connected field comprising four major clusters: physiology, the mechanics of research, subjects of field research, and lab research, survival and outcomes. In the large 'species' cluster, several terms occur (eg. systematics, taxonomy, and phylogeny) which the ANZSRC classifies in the FoR code Evolutionary Biology (0603). In addition, 'fossil record' (66 times) and 'fossil' (82 times) occur in the 'species' cluster. These terms are associated with palaeontology (an exclusion from Zoology) and are classified by the code 0403 for Geology. A close relationship is also evident with ecology (FoR code 0602) in the occurrence of the terms 'conservation concern' (68), 'conservation status'

(104) and 'agroecosystem' (76) in the 'bird' dominated cluster. Animal physiology is excluded from Zoology by the ANZSRC, belonging instead to Physiology (0606). Although the results are inconclusive, the occurrence of terms relating to physiology (eg. gene, dna, kidneys, and gland) in the 'cell' cluster suggests that a substantial amount of content includes references to animal physiology.

Nursing 1110

The Nursing journals dataset consisted of 15,000 items, drawn from 64 journals. The threshold for term occurrence was set at 15, from which the VOSviewer software generated 1561 relevant terms. One of the five clusters, seen in Figure 3, suggests there is high level of content about nursing education. Visible terms on the map include: student, learning, lecturer, and academic. Other terms, not visible on the map but included in the cluster are: clinical placement, nurse education, teaching, and curriculum. The ANZSRC does not list nursing education as an exclusion from the Nursing FoR code, however within the code 1302 for Curriculum and Pedagogy is 130209 for Medicine, Nursing and Health Curriculum and Pedagogy. 'Public health' is among the terms extracted, as is 'nursing home resident', both of which are specified exclusions in the Nursing FoR.

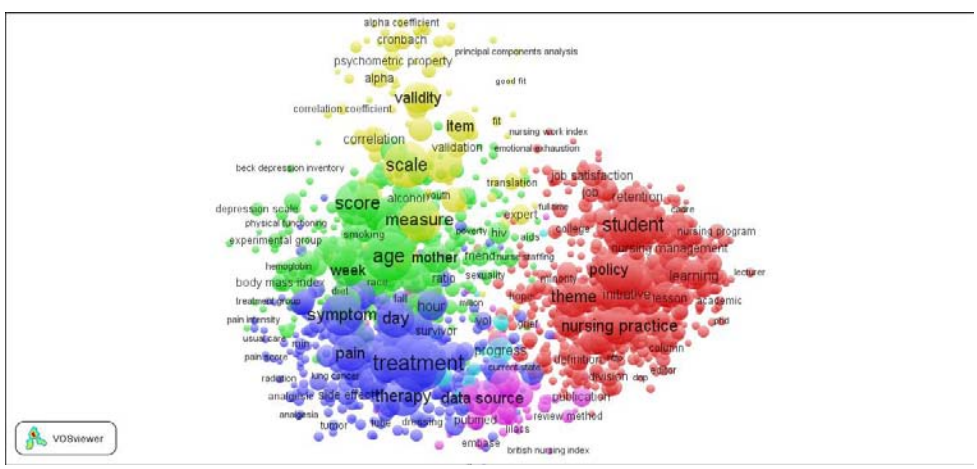


Figure 3. VOSviewer map of journal content for Nursing (FoR code 1110)

Pharmacology and Pharmaceutical Sciences 1115

The largest dataset, with 79,000 items, was the content of 152 journals from the Pharmacology and Pharmaceutical Sciences FoR. A threshold of 80 was used for term occurrences and 2,015 relevant terms were generated by VOSviewer. The map (Figure 4) displays three strong clusters, representing: aspects of trials; drug release; and discovery, as well as two smaller clusters relating to pharmacokinetics and drug responses. Medicinal chemistry, which includes

proteins and peptides, is a stated exclusion from the 1115 code, yet the term protein occurs over 6,000 times and is central to the cluster dominated by 'cell'. Proteins are, however, an important aspect of pharmacology and no conclusion can be reached as to whether the journals' content is focussed on pharmacology or extends into Medicinal Chemistry (FoR code 0304).

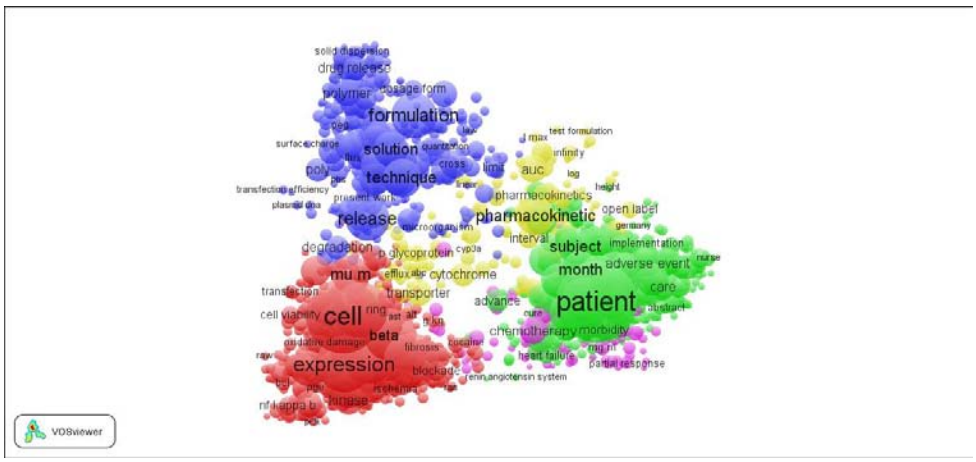


Figure 4. VOSviewer map of journal content for Pharmacology and Pharmaceutical Sciences (FoR code 1115)

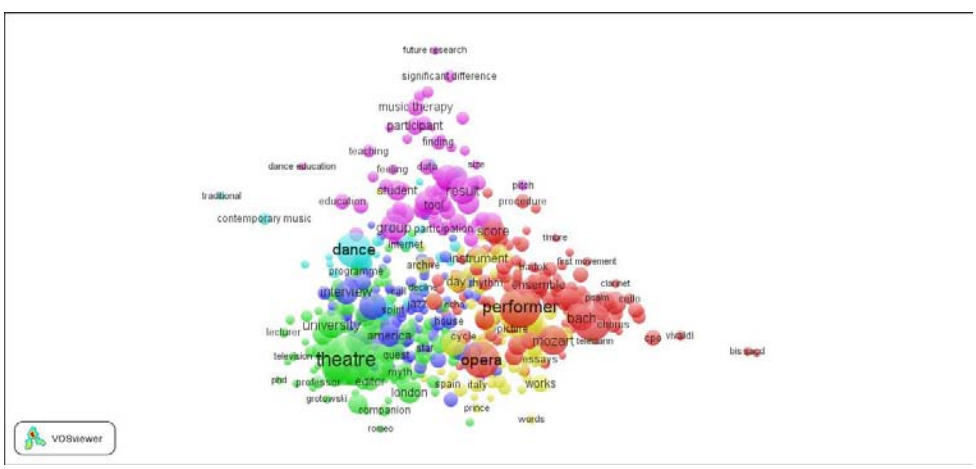


Figure 5. VOSviewer map of journal content for Performing Arts and Creative Writing (FoR code 1904)

Performing Arts and Creative Writing 1904

A threshold of 10 term occurrences was used to create the map (Figure 5) for the Performing Arts and Creative Writing journal set, which consisted of 9,000 items from 59 journals. Even with the low threshold, only 594 relevant terms were

generated by the software. Less defined clusters are evident in the map, which shows five large and one smaller cluster, with extensive scattering and overlap. Classical music is represented by a strong cluster, as is theatre. The term 'theatre' occurs 418 times, far exceeding any other terms. One cluster relates to education and research, including the terms: dance education; education; pedagogy and music education. Combined, these terms occur over 120 times. The ANZSRC classifies 'Creative Arts, Media and Communication Curriculum and Pedagogy' in FoR 1302, however, like Nursing, education is not a stipulated exclusion from the 1904 code. Creative writing is evident in the theatre cluster, but does not appear to be represented by a cluster of its own. 'Biography' (67 occurrences), which is specifically listed in 2103 Historical Studies, appears in the classical music cluster. With fewer occurrences, 'television', 'cinema' and 'film' are all exclusions from 1904 and are classified with the 1902 code for Film, Television and Digital Media. The term 'visual art' (14 occurrences), which has a specified code in the ANZSRC, also appears in the map within the central, scattered modern music cluster.

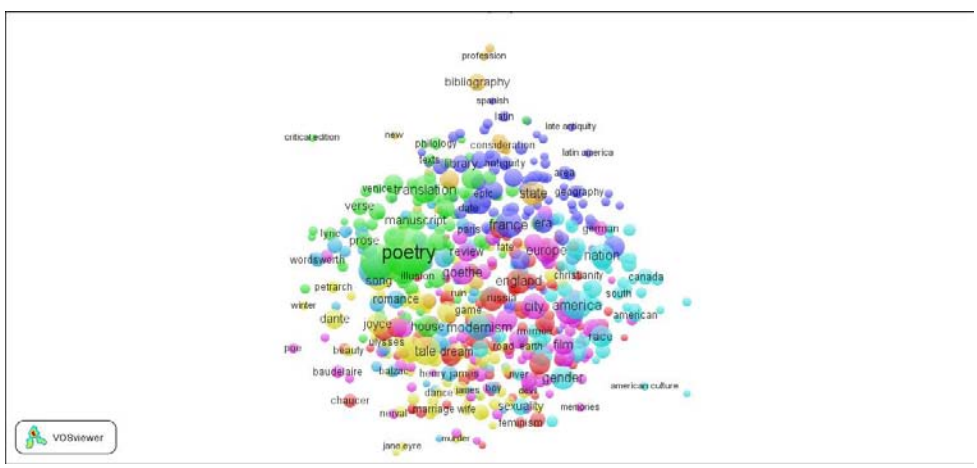


Figure 6. VOSviewer map of journal content for Literary Studies (FoR code 2005)

Literary Studies 2005

With 166 journals and over 30,000 items, the Literary Studies dataset was mapped using a threshold of 20 term occurrences. VOSviewer generated 589 relevant terms. The threshold was set lower than the similarly sized Zoology dataset (threshold of 25) to ensure a sufficient number of relevant terms were selected. Although several clusters can been seen in the map (Figure 6), extensive overlap exists between them. Poetry dominates in the most frequently occurring terms and is clearly depicted in a relatively defined cluster. Centrally located, but not visible, is 'theatre' which occurs 197 times. The ANZSRC does not stipulate 'theatre' as an exclusion from Literary Studies, however performing arts is classified by the FoR 1904. Similarly, 'film' and 'cinema' (occurring 155 and 109

times) are not stated exclusions, yet these terms are associated with the 1902 code. A stated exclusion is 'biography' (occurring 132 times in the poetry cluster), which should be classified under FoR 2103. Terms associated with education (classified by 1302), such as 'education', 'teacher' and 'student', also appear in the map, with over 220 occurrences combined.

Archaeology 2101

The Archaeology dataset was small, comprising 53 journals which produced over 5,000 items. For this reason a low threshold of 10 was set, with VOSviewer generating 583 terms as relevant to map (Figure 7). There was a high level of uniformity in occurrence of extracted terms; the highest being 157 followed by other terms occurring at regularly decreasing rates. Although some scatter of clusters is evident in the map, the field has four relatively defined clusters. These relate to burial sites and finds, representation and collections, administration and interpretation, and social and environmental aspects. Very few terms associated with other fields were extracted from the dataset. 'Anthropology' (classified by the code 1601), which occurred 27 times is a stipulated exclusion from Archaeology. 'Biography' (classified by the Historical Studies code 2103) occurred 21 times, and 'education' occurred 16 times.

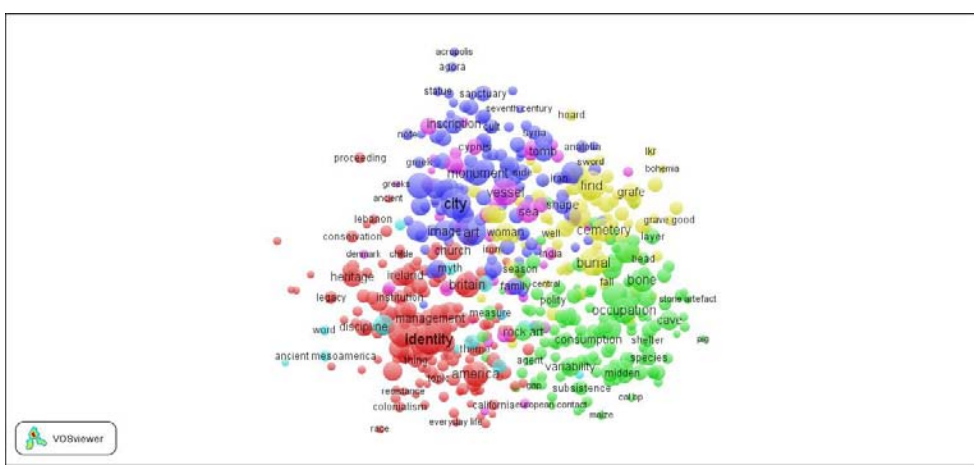


Figure 7. VOSviewer map of journal content for Archaeology (FoR code 2101)

Historical Studies 2103

The Historical Studies journals (180) produced a large dataset of over 48,000 items. A threshold of 25 was used to map the field, resulting in 1,050 relevant terms being generated. Despite the field's broad subject coverage, ANZSRC defines it as "history of peoples, nations or geographic regions", the map (Figure 8) shows four relatively strong clusters, with a fifth scattered cluster. The main clusters represent: war; American history; sources, interpretation, and social pursuits; and monarchy and church. The scattered cluster relates to colonisation

and migration. Several important overlaps are evident in the terms and partial clusters in the map. It is possible that these are due to sources consulted in historical research. For example, the ANZSRC stipulates that art history, history of architecture, and "history of specific concepts or fields of study which defy geographical classification" are excluded from the 2103 code. Each of these areas is represented by extracted terms. 'Art' occurs 588 times and 'architecture' 144 times. 'Environmental history' and 'philosophy' (classified by FoR 2202 'History and Philosophy of Specific Fields) occur 75 and 210 times, respectively. In addition, the occurrence of 'archaeology' (112), 'anthropology' (64), 'novel' (166), 'fiction' (171), 'poetry' (157), and 'shakespeare' (87) indicates there is substantial overlap with other FoR codes.

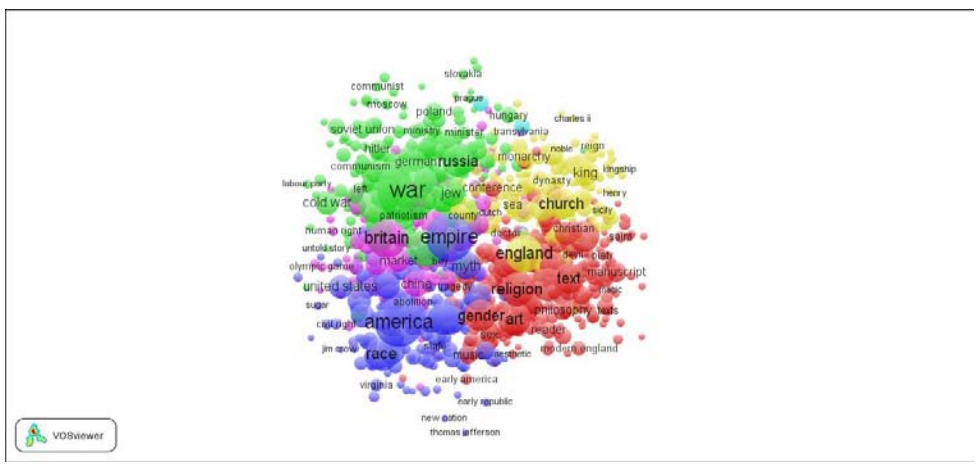


Figure 8. VOSviewer map of journal content for Historical Studies (FoR code 2103)

Overlapping Fields of Research codes

The matrix in Table 2, presents the findings for journal content that indicated overlapping FoR codes and includes the full range of codes identified in the analysis. Correspondence within a FoR code (eg. 0101 and 0101) is not displayed. A symbol in a grey cell indicates overlap, a large symbol in a black cell indicates a high level of overlap, either through clearly visible clusters of terms or a high occurrence of terms. For the latter, the number of occurrences of a term or similar terms combined, such as 'art' and 'artist', was calculated as a percentage of the highest term occurrence in the map. Terms that occurred more than 50% of the highest term's occurrence were deemed to illustrate stronger overlap. Strong overlaps can be observed in three of the science FoRs: Zoology, Nursing and Pharmacology. Journal content in the HSS fields presented a high degree of overlap with each other and with the education FoR, but there was not the same level of strength (assessed by term occurrence) as observed in the science fields.

Table 2. Matrix of overlap between Field of Research codes' journal sets

FoR	01 01	01 05	03 04	04 03	06 02	06 03	06 06	06 08	11 10	11 15	11 17	13 02	16 01	19 02	19 04	19 05	20 05	21 01	21 03	22 02
0101		x																		
05	x																			
0304										**X**										
0403								x												
0602								x												
03								**X**												
06								**X**												
08				x	x	**X**	**X**													
1110											x	**X**								
15			**X**																	
17									x											
1302									**X**						x		x	x		
1601																		x	x	
1902															x		x			
04												x		x		x	x		x	
05															x					
2005												x		x	x				x	
2101												x	x						x	
03													x		x		x	x		x
2202																			x	

Discussion

The analysis of terms extracted from the articles and titles in journals assigned a single FoR code indicates the assignment of one code is likely to be an inadequate representation of their content. In particular, overlaps with the education field were seen in a number of the journal sets. This is evident in the Nursing and, to a lesser extent, the Performing Arts and Creative Writing maps and matrix. Potentially, there are strong overlaps with other fields of research in Pharmacology and Zoology, but additional context for the frequently occurring terms is needed in order to reach conclusive findings.

It is interesting to note the difference in the interplay of FoR codes within science journal content compared with the content of HSS journals. Overall, there appeared to be fewer overlapping FoRs in the content of science journals, whereas the HSS display higher frequency of overlapping fields, with the exception of Archaeology which presented as a self-contained field. This may be a function of the language used in the fields, in that the sciences lexicon tends to be precise and based on well-defined terminology, while the language of HSS lacks the specificity of the sciences. Some of the overlaps observed in the HSS journal content were shared across the different fields, such as for the term 'biography' which occurred in all datasets despite being a stipulated exclusion from three of the FoR codes examined.

A classification scheme such as the ANZSRC is artificial by nature, whereas journal content comprises the natural language used by authors. This study did not seek to test the classification scheme per se. It instead sought to determine if its application to journals is likely to disadvantage individual researchers. As an exercise in the study of classification these observations are interesting, but transferred to research evaluation activities and the impact on researchers, they are more concerning. Article content with stipulated exclusions in a FoR has the potential to disconnect both the researcher and their organisational unit from the research output. For a researcher in a productive unit the impact of losing a few research outputs to another FoR code will be minor. For smaller and less productive units, the loss could mean their field is not assessed in the ERA if they do not meet the threshold required to make a submission. This scenario is possible in all the fields analysed in this study and is a direct result of the FoR code mechanisms used to create a structure for the ERA. Introduced relatively recently (2009), the ERA's use of FoR codes to classify journal article outputs has not been explored widely and the full impact this mechanism is yet to be realised. However, research to date (Bennett, Genoni & Haddow, 2011; Kwok, 2013) suggests that the way in which FoR codes are applied in the Australian research assessment model is flawed.

While there is correspondence between the FoR code assigned to the journals examined in this study and their content, there is also substantial overlap with other fields. The evidence of multidisciplinary research found in the journal sets suggests the assignment of FoR codes, particularly a single code, is not an accurate or consistent method to classify article content. Further research is needed to understand the implications of an overarching classification scheme in research evaluation and several aspects of this study raised questions that could be pursued, such as the effect of a field's language use on the selection of relevant terms from abstract and title data. Both the number of terms and the nature of terms extracted from the datasets suggest that future studies need to consider approaches that will enable a more precise understanding of the context of important terms, particularly in HSS. In this regard, a sensitivity analysis is anticipated for the full version of the current paper. It will involve multiple thresholds and creating sub-samples of the publications in each FoR.

Conclusion

When a researcher selects a journal for publishing they do so based on a number of factors relating to audience, perceptions of quality and publishing processes. The FoR classification of journals introduces another consideration that may have little relevance to these. However, in the ERA the assignment of FoR codes to journals is not an insignificant issue. Given the importance of aligning researchers' FoR codes with the publishing journals' codes, the overlap between fields identified in this study indicate there is the potential for researchers to be disadvantaged if their contribution is assessed and aligned with a different

organisational unit. In a higher education environment where accountability and value for money are imperatives at senior management and policy-making levels, a poor assessment (or no assessment) in the ERA can affect an organisational unit's survival.

The findings of this study suggest that the assignment of FoR codes to journals, while serving the purpose of effortless classification at the macro level, are inadequate for the purpose of valuing contribution at the micro level. If, as the ESF (2012) asserts, connecting information about a researcher to their outputs is one of the primary challenges of research evaluation, then the assignment of FoR codes to outputs and researchers in the ERA is failing to address this challenge.

References

Australian Bureau of Statistics. (2008). *Australian and New Zealand Standard Research Classification (ANZSRC)*. Retrieved November 6, 2012 from: http://www.abs.gov.au/ausstats/abs@.nsf/0/6BB427AB9696C225CA2574180004463E

Australian Research Council. (2012). *ERA 2012 Journal List*. Retrieved November 6, 2012 from: http://www.arc.gov.au/era/era_2012/era_journal_list.htm

Australian Research Council. (2011). *ERA 2012 Submission Guidelines*. Retrieved November 6, 2012 from: http://www.arc.gov.au/pdf/era12/ERA2012_SubmissionGuidelines.pdf

Bennett, D., Genoni, P. & Haddow, G. (2011). FoR codes pendulum: Publishing choices within Australian research assessment. *Australian Universities' Review*, 53, 2, 88-98.

Box, S. (2010), Performance-based funding for public research in tertiary education institutions: Country experiences. In OECD, *Performance-based Funding for Public Research in Tertiary Education Institutions: Workshop Proceedings*. OECD Publishing. http://dx.doi.org/10.1787/9789264094611-6-en

Butler, L. (2010). Impacts of performance-based research funding systems: A review of the concerns and the evidence. In OECD*, Performance-based Funding for Public Research in Tertiary Education Institutions: Workshop Proceedings*. OECD Publishing. http://dx.doi.org/10.1787/9789264094611-7-en

European Science Foundation. (2012). *Evaluation in Research and Research Funding Organisations: European Practices.* A report by the ESF Member Organisation Forum on Evaluation of Publicly Funded Research. Retrieved November 6, 2012 from: http://www.esf.org/index.php?eID=tx_nawsecuredl&u=0&file=fileadmin/be_user/CEO_Unit/MO_FORA/MOFORUM_Eval_PFR__II_/Publications/mof_evaluation_final.pdf&t=1357886978&hash=281487d267f2bbbcd1d07ad475d31a0567e58a79

European Science Foundation. (2011). *The Classification of Research Portfolios.* Member Organisation Forum on Publicly Funded Research, Working Group on "Comparative Research Portfolios". Retrieved November 6, 2012 from: http://www.esf.org/index.php?eID=tx_nawsecuredl&u=0&file=fileadmin/be_user/CEO_Unit/MO_FORA/MOFORUM_Eval_PFR__II_/3rd_Workshop/Classification.pdf&t=1357886978&hash=596e8e371eba8cb8cdfefcb998d550e9469b6493

Goldfinch, S. & Yamamoto, K. (2012). *Prometheus Assessed? Research Measurement, Peer Review and Citation Analysis*. Oxford: Chandos.

Higher Education Funding Council for England. (2005). *RAE 2008: Guidance on Submissions*. Retrieved November 6, 2012 from: http://www.rae.ac.uk/pubs/2005/03/rae0305.pdf

Kwok, J.T. (2013). *Impact of ERA Research Assessment on University Behaviour and their Staff*. NTEU National Policy and Research Unit. Retrieved April 24, 2013 from: http://www.erawatch.org.au

Moed, H.F. (2005). *Citation Analysis in Research Evaluation*. Dordrecht: Springer.

OECD. (2007). Revised Field of Science and Technology (FOS) Classification in the Frascati Manual. Retrieved November 6, 2012 from: http://www.oecd.org/science/innovationinsciencetechnologyandindustry/38235147.pdf

Tertiary Education Commission. (2012). *Performance-based Research Fund: Quality Evaluation Guidelines 2012*. Retrieved December 2, 2012 from: http://www.tec.govt.nz/Documents/Forms%20Templates%20and%20Guides/PBRF-2012-Guidelines-Sept12.pdf

Van Eck, N.J., Waltman, L., Noyons, E.C.M., & Buter, R.K. (2010). Automatic term identification for bibliometric mapping. *Scientometrics*, 82(3), 581-596

Van Eck, N.J., & Waltman, L. (2010). Software survey: VOSviewer, a computer program for bibliometric mapping. *Scientometrics*, 84(2), 523-538

Waltman, L., Van Eck, N.J., & Noyons, E.C.M. (2010). A unified approach to mapping and clustering of bibliometric networks. *Journal of Informetrics*, 4(4), 629-635

MODEL TO SUPPORT THE INFORMATION RETRIEVAL PROCESS OF THE SCIENTIFIC PRODUCTION AT DEPARTMENTAL-LEVEL OR FACULTY-LEVEL OF UNIVERSITIES

Víctor Bucheli[1], Juan Pablo Calderón[2], Fabio González[3], Bopaya Bidanda[4], Juan Alejandro Valdivia[4] and Roberto Zarama[5]

1 vbucheli@uniandes.edu.co
Departamento de Ingeniería Industrial, Universidad de los Andes, Cr 1 No 18A-12, 111711, Bogotá (Colombia) - Ceiba, Complex Systems Research Center, Bogotá (Colombia)

2 ju-cald1@uniandes.edu.co
Departamento de Ingeniería Industrial, Universidad de los Andes, Cr 1 No 18A-12, 111711, Bogotá (Colombia) - Ceiba, Complex Systems Research Center, Bogotá (Colombia)

3 fagonzalezo@unal.edu.co
MindLab, Departamento de Ingeniería de Sistemas e Industrial, Universidad Nacional de Colombia, Bogotá (Colombia).

4 bidanda@engr.pitt.edu
Department of Industrial Engineering, University of Pittsburgh, Pittsburgh, PA 15261, (USA).

5 alejo@macul.ciencias.uchile.cl
Departamento de Física, Facultad de Ciencias, Universidad de Chile.
Ceiba, Complex Systems Research Center, Bogotá (Colombia)

6 rzarama@uniandes.edu.co
Departamento de Ingeniería Industrial, Universidad de los Andes, Cr 1 No 18A-12, 111711, Bogotá (Colombia) - Ceiba, Complex Systems Research Center, Bogotá (Colombia)

Abstract
Bibliographic databases such as Thomson Reuters' Web of Science (WoS) or Elsevier's Scopus support search filtering by country or institution. However, the study of the scientific production at internal levels of organizations (universities) such as departments or faculties is error prone. In this paper, it shows common errors to retrieve papers in WoS at departmental-level or at faculty-level. We propose a method to support the information retrieval process at internal level of universities. The method is composed by an exhaustive search strategy and a Bayesian model to estimate the attribution of universities' papers that belong to a given department or faculty. The method was validated on two real cases with promising results. This work is a research in progress; the contrast with other

methods and other cases of evaluation are proposed as future work. Nevertheless, it could open new opportunities to scientometric studies and research policy.

Conference Topic

Management and Measurement of Bibliometric Data within Scientific Organizations (Topic 9) and Old and New Data Sources for Scientometric Studies: Coverage, Accuracy and Reliability (Topic 2).

Introduction

Literature reports that indexed documents between 1980 and 2009 in systems such as WoS Web of Science —Thomson Reuters system— or Scopus —Elsevier system— exceeds 32 millon (WoS) and 37 millon (Scopus) publications, respectively (Jacsó P., 2009). The large data sizes of publication databases illustrate the information overload in academic organizations (Allen & T. D. Wilson 2003); in addition, the papers retrieved from the listed systems are the input to scientometrics studies, research performance evaluation, and research policy. These bibliographic databases search filtering by country or institution. Garg K.C. (2003) shows that several number of studies have drawn from these new disciplines to evaluate scientific activity by country, research area, and institution. Furthermore, the study of the scientific production at internal levels of organizations (universities) is not efficient; in addition, the information retrieval process is identity uncertainty—the identity uncertainty refers to how the objects in a data base are not labelled with unique identifiers (Hanna P., 2003).

The process of information retrieval from publications databases for a specific department or faculty is ambiguous. This is because of name variations: a department or a faculty may have multiple names and multiple departments or faculties may share the same name. Such name ambiguity affects the performance of document retrieval and may cause improper document attribution to departments or faculties. In addition, due to name misspellings, translation, transliteration, and inconsistent inclusion of initials and pseudonyms affect the performance of document retrieval process. These factors affect the correct department-level or faculty-level attribution and it is a challenge for bibliometric and scientometric analysis. Thus, the incorrect document attribution to internal entities of universities is not conducive to document retrieval of a single department or faculty.

We propose a method based on a priori information about the previous intellectual production of department or faculty. It is used to develop a search strategy and a Bayesian model. The proposed search strategy is exhaustive; it means, it retrieves full and relevant publications. To improve the scalability, accuracy, precision, and recall of the search strategy, we develop a classification model (probabilistic Bayesian model-PBM) that estimates the probability of a given document belonging to a particular department or faculty. Using one specific case, The

Department of Industrial Engineering of the University of Pittsburgh (DUP), we report the data search strategy and the classifier PBM. We describe the quality of data obtained through machine learning standard performance measurements.

This paper is organized as follows, section two presents a brief literature review and the problem of retrieving the intellectual production of department-level or faculty-level. Section three outlines the method to support the document retrieval process. Section four highlights the results of the method applied on: Department of Industrial Engineering – University of Pittsburgh (DUP) and Faculty of Engineering – Universidad de los Andes (Colombia) (FUA). Finally, in the last section, we discuss possible applications and future work.

2. The intellectual production of departments or faculties.

2.1. Related work.

The name disambiguation methods are one approach to support the document (scientific publications) retrieval process at departmental-level or faculty-level of universities. Specifically, the authorship uncertainty and the affiliation disambiguation are challenges for Bibliometrics and Scientometrics arenas. It is a problem since 1969 (Garfield E., 1969) and author name disambiguation within bibliographic databases is a very active research area in computer science. Here, disambiguation approaches are based on machine learning paradigms, for a review of author name disambiguation procedures and algorithms, see (Smalheiser N. & Torvik V., 2009; Tang L. & Walsh J. P., 2010; Koppel M. & Schler J., 2009 and Stamatatos E., 2009). In the last two years, new procedures and algorithms of author disambiguation have been proposed (Gurney T., Horlings E. & Besselaar P., 2012; Morillo F. & Santabárbara I., 2013; Jiang W. & Ding X., 2013) .

On the other hand, universities could have information systems to track their publication outputs or information retrieval systems such as ResearcherID. However, researchers could not input all relevant information on a timely basis or they could enter erroneous information, which end up corrupting restricted data sets. Thus, these data sets need to be constantly checked and refined. Dervos et all (2006) show a pilot version of the Universal Author Identifier system, codenamed UAI_Sys and describe the critical non functional requirements, for example: user authorization policies, flexibility, durability and ensuring the data integrity. Smalheiser N. and Torvik V. (2009) show that this solution can not be reliable.

Bibliometric studies about the affiliation disambiguation in WoS say: the retrieves information on affiliation is diverse and it may generate some degree of uncertainty. (García-Zorita C. et all, 2005). Few studies have focused on institutional affiliation, where the problems encountered are related to the lack of standardization in the institutional addresses of author affiliation (Hood W &

Wilson C, 2003). In general, the need for standards in scientometrics studies has been reported in the literature (Glänzel W., 1996; Raan A.F.,1997; Hood & Wilson 2003). Jiang Yong, et all (2011) present a clustering method based on normalized compression distance for the purpose of affiliation disambiguation.

The method exposed in this paper was tested in WoS and it allowed us to compare and benchmark the productivity by departments or faculties of different universities.

2.2. The document retrieval process of intellectual production of departments or faculties.

The information and indicators about scientific publications are basic elements for knowledge management (Phillipswren G. & Forgionne G., 2006; Allen D. & Wilson T. D., 2003). The universities have gained interest in this kind of information. This has become more relevant because the number of publications and citations are inputs of the academic rankings, the allocation of resources, and scientific recognition (Geiger 2004; Enserink 2007).

We describe a data search strategy for retrieve the intellectual production of a single department-level or faculty-level. It is delineated by filtering problems in WoS—it can be extended to other systems such as Scopus or Scholar google. In WoS, the field tags AD Address, OG Organization and SG sub-organization are used to retrieve documents related to one organization: department-level or faculty-level. These are defined as searches for the institution and/or place names in the Addresses field within a record. (Thomson Reuters, 2011).

In this paper, we propose a document retrieval method to identify the documents that belong to a single academic organization within universities. The correct attribution represents an unsolved problem for information science. This is due to name variations; for instance, departments or faculties may have multiple names and multiple departments or faculties may share the same name. The name ambiguity affects the performance of document retrieval and may cause improper document attribution to departments or faculties. In addition, the name misspellings, translation, transliteration, and inconsistent inclusion of initials and pseudonyms affect the performance of document retrieval process. The proposed method take into account these issues, however, other cases are not consider, such as joint appointment affiliation.

2.2.1. Common errors.

In this work, we build search statements with the following field tags: AD and OG and non-relevant documents are retrieved, for instance, we search the publications of La Universidad Nacional de Colombia from WoS and this university appears with multiple name labels: *univ nacl colombia, natl univ colombia, univ nacl or natl univ*. In other case, The Department of Industrial

Engineering of the University of Pittsburgh appears as: Ind Engn Dept, Pittsburgh; Dept Ind Engn, Pittsburgh; or Univ Pittsburgh, Swanson Sch Engn, Dept Ind Engn, Pittsburgh. On the other hand, The University of Pittsburgh and the city of Pittsburgh share the same name label in WoS. We use the search statement ad=("univ pittsburgh"), and it retrieves documents from both name labels: the university of Pittsburgh and the Carnegie Mellon University allocated in Pittsburgh—labelled as Carnegie Mellon Univ, Pittsburgh. In the case OG=("univ pittsburgh"), this also retrieves non-relevant documents.

2. Description of a method to retrieve the intellectual production

In order to illustrate the proposed method, we use The Industrial Engineering Department of the University of Pittsburgh, see Appendix 1.

2.1 Data search strategy to retrieve the intellectual production

In order to retrieve the corpus data, we search the documents published by university— WoS search strategy ad=(university name). According to the supervised learning methods, the corpus data will split in two data sets: training data set and test data set. Here, the data search strategy is the base line to construct the training data set, thus, the data search strategy is composed of two steps.

First, the initial search strategy is configured based on a priori information about the intellectual production of department or faculty and it is characterized into four sets: staff, journals, socio-semantic and explicit information about the academic unit in the address field. Second, we retrieve a list of papers based on the initial search strategy. Here, each document is classified as relevant and non-relevant depending if the paper belongs to the faculty (or department) or not. This list and the classification are validated by an expert group; a professor committee or the dean of the academic unit. Their recommendations and suggestions are taken into account and they are integrated into the new search strategy. Additional restrictions to filter out documents are also integrated.

In the following section we explain the configuration of the initial search strategy and the final search strategy. We use boolean notation to explain the configuration of the data search strategies. Finally, the PBM is presented.

2.1.1. Configuration of the initial search strategy

The search strategy retrieves the publications of the university and it construct the corpus data in the proposed method; thus, the university set **(U)** is the group of documents in which the name of the university appears in the address of at least one of the authors. The initial set **(I)** or initial search strategy is composed by the union of the following groups: staff set **(S)**, socio-semantic network set **(O)**, journals set **(J)**, and the name of the internal unit set **(A)**. Hence, **I = S** OR **O** OR **J** OR **A**.

The staff set **(S)** is related to the documents authored by the faculty members. In this way, the set **(S)** is the group of documents written by one member of the faculty, whereby the name of the university appears in the institutional affiliation address.

The socio-semantic set **(O)**: in this context represents the combination of semantic and author information—Roth C. and Cointet J. (2010) present a similar representation. In this paper, the socio-semantic information represents a network where concepts and authors appears together. Each document is represented by a specific semantic category. Thus, set **(C)** contains the group of documents in which one concept appears in the title or abstract. The list of concepts can be provided by the academic unit or extracted automatically from previous publications. In this work, we process title and abstract textual information and build up a subset of n-grams[115] (Croft 2010). Then, we automatically create a concept lists. On the other hand, the set **(O)** is also related to the documents authored by the faculty members, hence, **O**=**C** AND **S**.

The journals, set **(J),** is the group of documents published in the same journals as previous publications by the faculty, in which the name of the university appears in the address. We build a list of journals in which the unit has published before and this is the input to build the set **J**.

The explicit academic unit, set **(A),** is the group of documents in which the name of the academic unit appears in the address of one of the authors; for instance, the engineering school refers to as ENG, the physics schools refers to as PHY.

2.1.2 Expert validation

The search statement for the initial set is applied to retrieve a list of papers. Each paper is validated by experts as explained above, then, the non-relevant papers are removed from the list and a new list of the paper's titles of non-relevant documents are included into the restriction, set **(T)**. This is used as input for the new search strategy. We use another group of restrictions in the address field, set **(Q)**. This set is similar to set **(A)**, but it contains other academic units that not have been taken into account. The final search strategy or retrieved set **(R)** is the group of documents of **(I)** that do not belong to sets **(T)** and **(Q)**. The intellectual production of department-level or faculty level is retrieved through the final search strategy or retrieved set **(R)**. Then, **R = I** NOT (**T** OR **Q**).

2.1.3 A classification model for academic units tracking

The information retrieved by the final search strategy is the input to the Probabilistic Bayesian Model (PBM), which is based on a Naive Bayesian Model. The information retrieved can be separated into relevant and non-relevant. Thus,

115 We used Tinasoft software http://tina.csregistry.org/

the information retrieval methods and classifier models have a similar purpose (Croft 2010).

The proposed classifier is Naive Bayes, which follows the Bayesian Theorem. The basic assumption of the Naive Bayesian Model, called class conditional independence, is the independence of features. It is made to simplify the computation (it is considered to be naive). This kind of classifier is commonly used in spam classification, email sorting, routing filtering, and document classification (Baeza-Yates R., 1999; Manning C., 2008).

The aim of this model is to classify the university documents into two classes: documents that belong to the department-level or faculty-level (relevant) and documents that do not belong (non-relevant). The variables are related to each set defined in the search strategy: staff, journal, socio-semantic, explicit address and class (relevant and non-relevant). We built a corpus data with the final search strategy of the faculty and each document is marked as relevant or non-relevant. The complete corpus data is split into two sets (training data set and test data set). The Probabilistic Bayesian Model (PBM) was build with the first data set (training set). We use Weka cross-validation 10 folds and percentage split to train the model (Hall M, et all 2009). This set is used to estimates the posterior probability of a new document belonging to the given unit as a function of the four variables listed below. We defined the model as:

$$p\,(R|J,S,O,A) = \frac{p\,(R\,)p\,(J,S,O,A|R)}{p\,(J,S,O,A)} \qquad (1)$$

To simplify the computations:

$$p\,(J,S,O,A|R) = p\,(J|R)\ \times p\,(S|R)\ \times p\,(O|R)\ \times p\,(A|R) \qquad (2)$$

This probability model is trained through a supervised learning method (training data set). The m-stimate (equation 3) is a method to estimate the probability of a new paper belonging to the relevant category; for instance, the probability of J given that R:

$$p(J|R) = \frac{n_j + p}{n_r + 1} \qquad (3)$$

Where, n_j is the number of relevant documents in J; n_r is the number of relevant documents; and p is a priori estimate for p(R|J). With this information, for each document defined as (4), the new paper is classified in the respective set given that (5).

$$d_i(j_i, s_i, o_i, a_i) \quad (4)$$

$$p(R)\, p(j_i|R) \times p(s_i|R) \times p(o_i|R) \times p(a_i|R)) > 0.5 \quad (5)$$

3. Experimental evaluation and results

The proposed data search strategy had better performance than the common search statements (AD or OG in WoS), in the case of DUP, the retrieve records were from 117 to 145, these values are taken into account as base line to evaluate the retrieval performance, as well as, standard measurements: error instances classified, precision, recall and the area under the receiver operating characteristic curve (ROC) (Witten I., 2005; Baeza-Yates R., 1999).

3.1. Experimental evaluation

We apply the proposed search strategy for DUP and 173 documents were retrieved. This highlight shows the correct attribution and how the proposed method support the document retrieval process at faculty-level. The proposed method retrieves more publications than the common search statements. The common search statements presented an average of lost of papers from 24% to 29%. Thus, the proposed search strategy is exhaustive. These results has been confirmed with the Faculty of Engineering – Universidad de los Andes (Colombia) (FUA) and the method is consistent.

3.2 Results

Table 1. shows the results in the case of Department of Industrial Engineering – University of Pittsburgh and Faculty of Engineering – Universidad de los Andes (Colombia). This shows the performance and the predictiveness of the model.

Table 1. Performance measurements of the proposed model.

	Instances wrongly classified	*Precision*	*Recall*	*Area under the ROC*
Department of Industrial Engineering – University of Pittsburgh	0.16%	0.997	0.494	0.984
Faculty of Engineering – Universidad de los Andes (Colombia)	0.48%	0.954	0.992	0.965

4. Discussion and future work

The identity uncertainty in the paper's institutional affiliation affects the performance of document retrieval process at department-level or faculty-level and may cause improper document attribution to departments or faculties.

We report a method to support the information retrieval process of the scientific production at departmental-level or faculty-level. We describe the quality of data obtained through standard performance measurements. The results show that it can effectively identify the correct attribution. This paper is a research in progress; as future work, the contrast with other methods and other cases of evaluation will be considered.

In order to apply the proposed method the search strategies and the PBM can be used as an integrated system, where, the search strategy can be implemented through RSS alerts (Rich Site Summary) from WoS and the RSS retrieves the publications periodically. On the other hand, the PBM can be implemented in Weka (Java application), which allows an automatic classification of departmental or faculty documents.

The proposed method allows to develop scientometrics studies at the different levels within universities and to compare and benchmark the research productivity by departments or faculties of different universities. The presented method shows the potential to evaluate and track knowledge production and support the research policies within universities.

As future work, this method can be extended to include research groups or research networks, as well as, other systems such as Scopus or Google Scholar.

References

Allen, D. & Wilson, T.D., 2003. Information overload: context and causes. New Review of Information Behaviour Research, 4(1), pp.31-44.

Baeza-Yates, R., 1999. Modern information retrieval, New York ;Harlow England: ACM Press ;;Addison-Wesley c1999.

Croft, W., 2010. Search engines : information retrieval in practice, Boston: Addison-Wesley.

Dervos, D. A., Samaras, N., Evangelidis, G., Hyvärinen, J., and Asmanidis, Y., 2006. The Universal Author Identifier System (UAI_Sys). Proceedings of the 1st International Scientific Conference, eRA: The Contribution of Information Technology in Science, Economy, Society and Education. Retrieved January 12, 2008, from dlist.sir.arizona.edu/1716

Enserink, M., 2007. EDUCATION: Who Ranks the University Rankers? Science, 317(5841), pp.1026-1028.

García Zorita, C., Martín Moreno, C., Lascurain Sánchez, M. L., & Sanz Casado, E., 2006. Institutional addresses in the Web of Science: the effects on scientific evaluation. Journal of Information Science, 32(4), pp. 378–383.

Garfield, E., 1969. British quest for uniqueness versus American egocentrism. Nature, 223(5207), pp. 763.

Garg K.C., 2003. An overview of cross-national, national, and institutional assessment as reflected in the international journal Scientometrics, Scientometrics, 56 (2), pp.169– 99.
Geiger, R., 2004. Knowledge and money : research universities and the paradox of the marketplace, Stanford Calif. Stanford University Press.
Glänzel, W., 1996. The need for standards in bibliometric research and technology. Scientometrics, 35(2), pp.167-176.
Gurney T., Horlings E., and Besselaar P, 2012. Author disambiguation using multi-aspect similarity indicators. Scientometrics, 91(2),pp. 435-449
Hall M, et al, 2009, The WEKA Data Mining Software: An Update; SIGKDD Explorations, Volume 11, Issue 1.
Hanna P., et al., 2003, Identity Uncertainty and Citation Matching, In NIPS, MIT Press.
Hood, W.W. & Wilson, C.S., 2003. Informetric studies using databases: Opportunities and challenges. Scientometrics, 58(3), pp.587-608.
Jiang Wu & Xiu-Hao D., 2013. Author name disambiguation in scientific collaboration and mobility cases. Scientometrics, pp.1-15.
Koppel, M., Schler, J. and Argamon, S., 2009. Computational Methods in Authorship Attribution, JASIST, 60 (1), pp. 9-26
Jacsó, P., 2009.Errors of omission and their implications for computing scientometric measures in evaluating the publishing productivity and impact of countries. Online Information Review, 33(2), pp.376-385.
Jiang, Y., Zheng, H. T., Wang, X., Lu, B., & Kaihua, Wu., 2011. Affiliation disambiguation for constructing semantic digital libraries. Journal of the American Society for Information Science and Technology, 62(6), 1029–1041.
Manning, C., 2008. Introduction to information retrieval, New York: Cambridge University Press.
Morillo F. & Santabárbara I. and Aparicio, J., 2013. The automatic normalization challenge: detailed addresses identification. Scientometrics, 91(1) pp.1-14 ,
Phillipswren, G. & Forgionne, G., 2006. Aided search strategy enabled by decision support. Information Processing & Management, 42(2), pp.503-518.
Raan, A.F.J., 1997. Scientometrics: State-of-the-art. Scientometrics, 38(1), pp.205-218.
Roth, C. & Cointet, J.-P., 2010. Social and semantic coevolution in knowledge networks. Social Networks, 32(1), pp.16-29.
Smalheiser NR, Torvik VI. Author name disambiguation. In: Cronin B, editor. Annual Review of Information Science and Technology. Vol. 43. 2009. pp. 287–313.
Stamatatos E., 2009. A survey of modern authorship attribution methods. Journal of the American Society for Information Science and Technology, 60 (3), pp. 538-556.
Tang, L et. Al, 2010. Bibliometric fingerprints: name disambiguation based on approximate structure equivalence of cognitive maps, Scientometrics, 84 (3), pp.763-784.

Thomson Reuters, 2011. Web of Knowledge - Science - Thomson Reuters. Web of knowledge. Available at: http://www.isiwebofknowledge.com/

Witten, I., 2005. Data mining : practical machine learning tools and techniques 2nd ed., Amsterdam ;;Boston MA: Morgan Kaufman.

Appendix 1.

Table 1: Data search strategy for the Department of Industrial Engineering of the University of Pittsburgh

SET	Search statement	Number of documents
U	AD=(UNIV PITTSBURGH) OR OG=(UNIV PITTSBURGH)	61,57
S	au=(Shuman,L*) OR au=(Bidanda,B) OR au=(Rajgopal,J*) OR au=(NORman,BA) OR au=(Besterfield-Sacre,M*) OR au=(Kharoufeh,JP) OR au=(Maillart,L*) OR au=(Prokopyev,O*) OR au=(Shankar,MR) OR au=(Schaefer,A*) not au=(Schaefer, at) AND (ad=(UNIV PITTSBURGH) OR og=(UNIV PITTSBURGH))	117
C	ts=(suite of test problems) OR ts=(severe plastic deformation spd) OR ts=(solid oxide fuel cell) OR ts=(cross wedge rolling cwr) OR ts=(behavior in electricity markets) OR ts=(stage liver disease) OR ts=(strategic behavior in electricity) OR ts=(leading cause of death) OR ts=(unequal area facility) OR ts=(several oligopoly models) OR ts=(set ofintuitive conditions) OR ts=(wedge rolling cwr) OR ts=(objective tabu search) OR ts=(expected lifetime orquality) OR ts=(power generating system) OR ts=(imposition of knowledge) OR ts=(optimal policy) OR ts=(mathematical programming) OR ts=(production costs) OR ts=(decisi on process) OR ts=(crew schedules) OR ts=(decision maker) OR ts=(control limit) OR ts=(natural history) OR ts=(stage liver) OR ts=(integer programming) OR ts=(genetic algorithm) OR ts=(outcome measures) OR ts=(generating system) OR ts=(numerical example) OR ts=(planned cost) OR ts=(several oligopoly) OR ts=(expected profit) OR ts=(decision problem) OR ts=(welded structure) OR ts=(process planning) OR ts=(duration curve) OR ts=(generating unit) OR ts=(product design) OR ts=(auditing practices) OR ts=(leading cause) OR ts=(engineering education) OR ts=(sofa scores) OR ts=(crew schedule) OR ts=(ofintuitive conditions) OR ts=(strategic behavior) OR ts=(expected lifetime) OR ts=(stochastic model) OR ts=(petri nets) OR ts=(engineering programs) OR ts=(plane strain) OR ts=(job rotation) OR ts=(waiting list) OR ts=(process mdp) OR ts=(robot arc) OR ts=(optimization problem) OR ts=(intellectual property) OR ts=(handling costs) OR ts=(rescheduling problem) OR ts=(alternative routings) OR ts=(planning process) OR ts=(institutional changeand) OR ts=(hollow shafts)	>100.000
O	**C** AND **S**	87
J	so=(IIE TRANSACTIONS) OR so=(EUROPEAN JOURNAL OF OPERATIONAL RESEARCH) OR so=(INTERNATIONAL JOURNAL OF PRODUCTION RESEARCH) OR so=(IEEE TRANSACTIONS ON POWER SYSTEMS) OR so=(JOURNAL OF ENGINEERING EDUCATION) OR so=(MEDICAL DECISION MAKING) OR so=(OPERATIONS RESEARCH) OR so=(ANNALS OF OPERATIONS RESEARCH) OR so=(NAVAL RESEARCH LogISTICS) OR so=(MANAGEMENT SCIENCE) OR so=(OPERATIONS RESEARCH LETTERS) AND (ad=(UNIV PITTSBURGH) OR og=(UNIV PITTSBURGH))	132
A	ad=(UNIV PITTSBURGH SAME Dept Ind Engn) OR og=(UNIV PITTSBURGH same Dept Ind Engn)	145
I	**S** OR **O** OR **J** OR **A**	216
Q	(ad=(UNIV PITTSBURGH same Business*) OR og=(UNIV PITTSBURGH same Business*))	295
T	The recommendations and suggestions of experts are taken account in a restriction set **T**	17
R	**I** NOT (**T** OR **Q**)	**173**

MOST BORROWED IS MOST CITED? LIBRARY LOAN STATISTICS AS A PROXY FOR MONOGRAPH SELECTION IN CITATION INDEXES (RIP)

Álvaro Cabezas-Clavijo[1], Nicolás Robinson-García[1], Daniel Torres-Salinas[2], Evaristo Jiménez-Contreras[1], Thomas Mikulka[3], Christian Gumpenberger[3], Ambros Wernisch[3]& Juan Gorraiz[3]

[1] *acabezasclavijo@gmail.com, {elrobin, evaristo} @ugr.es*
EC3: Evaluación de la Ciencia y de la Comunicación Científica, Departamento de Información y Comunicación, Universidad de Granada, Campus de Cartuja s/n E-18071 Granada (Spain)

[2] *torressalinas@gmail.com*
EC3: Evaluación de la Ciencia y de la Comunicación Científica, Centro de Investigación Médica Aplicada, Universidad de Navarra, Pamplona (Spain)

[3] *{christian.gumpenberger, ambros.wernisch, thomas.mikulka, juan.gorraiz} @univie.ac.at*
University of Vienna, Vienna University Library, Boltzmanngasse 5, A-1090 Vienna (Austria)

Abstract
This study aims to analyse whether library loan statistics can be used as a measure of monograph use and as a selection criterion for inclusion in citation indexes. For this, we conducted an exploratory study based on loan data (1000 most borrowed monographs) from two non-Anglo-Saxon European university libraries (Granada and Vienna) with strong social sciences and humanities components. Loans to scientists only were also analysed at the University of Vienna. Furthermore, citation counts for the 100 most borrowed scientific monographs (SM) and textbooks or manuals (MTB) were retrieved fromWeb of Science and Google Scholar.The results show considerable similarities in both libraries: the percentage of loans for books in national languages represents almost 96% of the total share and SM accounts only for 10%–13%.When considering loans to scientists only, the percentage of English books increases to 30%; the percentage of SM loans also increases (~ 80%). Furthermore, we found no significant correlations between loans and citations. Since loan statistics are currently insufficient for measuring the use of monographs, their suggested use as an applicable selection criterion for book citation indexes is not yet feasible. Data improvement and aggregation at different levels is a challenge for modern libraries in order to enable the exploitation of this invaluable information source for scientometric purposes.

Keywords
Loans, citation, usage metric, citation metric, monographs, books, book citation index

Conference Topics

Topic 1: Scientometric Indicators; Topic 2: Old and New Data Sources for Scientometric Studies

Introduction

Bibliometric indicators have increasingly been used for research assessment purposes since the 1970s. Among other uses, they have been applied to implement reward mechanisms within academia, exceeding their original purpose which was to serve as an aid for journal selection in university libraries. In this sense, bibliometric studies have rested primarily on two basic metrics: journals' impact factors and citations of papers. However, unlike journal articles for which new indicators have been developed in the last few years (SJR, SNIP, Eigenscore), monographs, an important scholarly communication channel for the humanities and social sciences fields, have been left behind. The absence of books in the main databases with bibliometric data has led evaluation agencies to consider monographs a minor scientific product. This has resulted, in many of these fields, in the devaluation of monographs. In fact, as shown in the UK, many researchers have shifted from books to journal articles as their preferred dissemination product due to the pressure exerted by national evaluations (Research Information Network, 2009).

There has been little exploration of usage indicators within the scientometric community as a proxy to measure the use and impact of academic materials. The advent of the digital format in academia has brought the development of new tools, such as journal hubs or repositories, which produce new use-derived indicators. These metrics represent a potential opportunity for applying alternative evaluation methods, aiming to complement or even replace the traditional bibliometric indicators based on citations. Projects such as COUNTER (Counting Online Usage of Networked Electronic Resources) or MESUR (Metrics from Scholarly Usage of Resources) have worked on this line of work, developing the necessary frameworks and standards to achieve such goals.

Likewise, the adoption of so-called web 2.0 tools by researchers has added a new dimension in which usage indicators can also be applied to measure the impact of research, not only within the academic community but also in society at large. The main characteristics of these indicators, known as Altmetrics (Priem et al, 2010), are: 1) they work at an item level, and 2) they can be obtained in real time. However, they also have many shortcomings, such as the evanescence of data or complexity in terms of apprehending their real meaning. The adoption of these indicators by journals such as PLoS One, publishing houses such as Nature Publishing Group, or major databases such as Scopus, highlight the level of acceptance they are gaining within the scientific community.

Despite the availability of e-books on scientific platforms and in library catalogues, these usage indicators do not perform well in this context such that they might be considered valuable alternatives for research assessment. However, there are other usage indicators that could be used instead. Library loans may capture the impact of books in ways that e-usage and Altmetrics cannot. This is particularly interesting in the fields of the social sciences and humanities, where monographs still play a significant role. This approach is not new as Price (1963) remarked that the "amount of usage provides a reasonable measure of the scientific importance of a journal or a man's work". Following this line of thought, loans could not only be used as proxies of dissemination, but also of relevance or importance, detecting books that may be considered top tier research outcomes for inclusion in citation indexes. This is as a relevant issue as evaluation agencies tend to assess based only the content of these indexes.

The launch of the Thomson Reuters' Book Citation Index (hereafter, BKCI) in 2010 introduced new elements of study for the bibliometric community. Beyond the citations that each book or book chapter gather from other citation indexes, this product also allows the analysis of publishers or citation patterns within book-oriented areas (Gorraiz et al, 2013; Leydesdorff & Felt, 2012; Torres-Salinas et al, 2012). Thomson Reuters states that "there is a need to select those publications that will most likely contain significant scholarship" and that "priority is given to books and book series that have relatively greater citation impact" (Testa, 2012). Such statements suggest that the mere indexing of monographs within this database is a sign of quality commensurate with that for journals and therefore it also suggests that solid methodology must be developed to ensure that the materials to be included in the future are chosen fairly. Currently, apart from certain bibliographic requirements and the citation impact, the company does not specify further criteria for inclusion.

Based on the premise that library statistics are an invaluable and underutilized source of information for assessment purposes, we explore in this paper the extent to which library loans might be used as a proxy for the measurement of monograph use. Furthermore, we test the feasibility of using library loans as a possible selection criterion for monographs in citation indexes.

Theoretical Background

The influence of monographs in the social sciences, especially in the arts and humanities, as the main communication channel between scholars has traditionally led to serious shortcomings when adopting bibliometric methodologies to analyse and assess research activity in these areas. Indeed, not only they lack sound and consolidated bibliometric measures, but their nature poses many limitations that must be overcome in order to apply the correct methodological procedure when employing them. In this sense, there are three main issues that should be resolved as a prior step before any kind of

methodological approach is taken: the definition of monograph types, the establishment of significant proxies of quality, and addressing differences across disciplines.

Regarding the first issue, monographs are extremely heterogeneous as content of a different nature can be found in them. According to Testa (2011), in the BKCI the following book formats are considered scholarly: dissertations, textbooks, books in series, reprinted/reissued content, translations and non-English content, biographies and reference books. However, this classification is of no use for bibliometric purposes as the ambiguity in the definition of each of these formats prevents us from establishing a relation between citation or usage patterns and document types as is done with journal articles (reviews, letters, notes, research articles, etc.). In relation to this, Torres-Salinas and Moed (2009) also point out the lack of book typologies as a shortcoming when assessing monographs and suggest three possible criteria: a) authorship of the monograph (authored versus edited works), b) research intensity (books primarily for teaching versus books primarily for research), and c) research focus (books for a specialized scientific audience versus books for broader audiences). However, such classifications cannot be found in databases or library catalogues, forcing us either to make no distinction, or use an erroneous classification which may lead to the wrong conclusions, or perform a manual classification. An alternative approach would be to classify them according to the traditional book categories described in library catalogues (i.e., handbook, manual, report and research work) and analyse them separately, deciding the level of research intensity after the analysis has been undertaken.

The second issue to take into account is the proxy chosen as a performance indicator when evaluating scholarly monographs. Until the launch of the BKCI, it remained extremely difficult to assess books using citation data for large sets of records; in fact, few studies can be found using large citation data sets regarding the evaluation of books (Gorraiz et al, 2012; Leydesdorff & Felt, 2012; Torres-Salinas et al, 2013). What is more, the few studies that use citations as a proxy warn that the patterns observed have many peculiarities that must be taken into account when they are used. One of them has to do with the aging of citations, as monographs seem to need much wider citation windows (Cronin, Snyder & Atkins, 1997) than two to five years, as is common with journal articles. This may be a good explanation for the findings described by Torres-Salinas et al (2012), who observed that more than 70% of the books and book chapters indexed since 2005 in the BKCI remained uncited.

In order to solve not only these shortcomings but also the issue regarding data availability, other proxies have been suggested in the literature for evaluating monographs based on: a) library holdings, such as the number of catalogue entries per book title in WorldCat® (Torres-Salinas & Moed, 2009), library bindings

(Linmans 2010), or even introducing an indicator of perceived cultural benefit (White et al,2009); b) document delivery requests (Gorraiz & Schlögl, 2006); c) publishers' prestige (Giménez-Toledo, Tejada-Artigas & Mañana-Rodríguez, 2012);d) book reviews (Zuccala & van Leeuwen, 2011). However, none of these has been adopted unanimously by the bibliometric community, mainly because obtaining the data is difficult and time consuming.

Finally, the last issue that needs to be resolved has to do with the different practices observed across disciplines. This is also observed when analysing journal articles, but it may be even more acute with monographs as these are most commonly used in the humanities and social sciences in which practices are more fragmented and there is a strong national factor biasing researchers' behaviour (Hicks, 1999). The role played by monographs is especially important in these fields as they form one of the main channels for scholarly communication (Hicks, 2004; Research Information Network, 2009, 2011; Williams et al, 2009).

Data & Methodology

We conducted a pilot study to test the extent to which library loans could be used as a proxy to measure monographs' relevance within the scientific community. Such analysis was performed in two non-Anglo-Saxon European university libraries: the library of the University of Granada (Spanish-speaking) and the Vienna University Library (German-speaking). Both of them are universities with several centuries of history and both libraries are universal with strong social sciences and humanities components.

Briefbackground of the institutions (structure) and description of their loan systems:

A) Granada:

One of the historic universities of Spain, the University of Granada was founded in 1531. Despite its encyclopaedic character, it is a university with strong social sciences and humanities components, these being areas to which 47.6% of the research staff are affiliated (University of Granada, 2011). Its library system comprises 21 libraries which are located within faculties, institutes and research centres, providing services to more than 80,000 students, 3,650 researchers and 2,000 technical and administrative staff. According to the last library report (University of Granada, nd), there are 1,042,575 monographs. The integrated library system developed by Innovative was established in 2001 in the library premises; the loan system was centralized by means of Millenium software and affords a number of different reports concerning loans and renewals of monographs and other materials.

B) Vienna

Founded in 1365, the Vienna University Library is the oldest university library in the German-speaking countries;[116] it is also the largest library in Austria with an inventory of over 6.8 million books. It comprises a main central lending library (2.6 million volumes) and 40 specialist libraries, providing researchers, teachers and students with specialist literature on almost all specific academic subjects. It provides services to approximately 91,000 students and a university staff of 9,400 employees, of whom 6,700 are academic. For some disciplines, such as medicine, veterinary medicine, economics, agriculture and technical subjects, there are additional universities with corresponding university libraries in the city. Since the winter semester of 1986, loans have been managed electronically; in 1989, an electronic catalogue was also introduced. In 1999, the Aleph integrated library system replaced the previous software. In 2010, around 4 million book loans and 65,000 active borrowers were counted.

Data retrieval and processing

Data were gathered from the Universities of Vienna and Granada library systems in December 2012 regarding loans from, respectively, 2001 and 2000 onwards. For every monograph copy we recorded the following fields: book title, author, location, publisher, country and year of edition, language, year of acquisition by the library, ISBN, number of loans, and number of renewals.

For the University of Granada, only copies of bibliographic material with more than 50 loans were recorded. Afterwards, loan data for every title were aggregated, regardless of their publication year or publisher. Books with the same title but different authors were also detected and considered separately. For the Vienna sample, loan data were analysed at the level of bibliographical records. Thus, all copies of a certain edition were automatically aggregated, but different editions of one work were covered separately. Also, Granada distinguished book types according to the library's classification. For this study, we decided to use three main book types:

- REF = reference books,such as dictionaries, etc.
- MTB = manuals, textbooks, handbooks, etc.
- SM= scientific monographs

These bibliographic types are assigned by librarians at the institution. However, some titles were assigned to two different types (MTB and SM). For those materials, all the copies were recoded as MTB. We also coded as MTB those monographs with the following words in their titles for Spanish and English language books: Course, Encyclopedia, Foundations, Introduction, Methods, Principles, Treaty, Grammar, Atlas, Compendium, Handbook and Textbook.

[116]http://bibliothek.univie.ac.at/english/about_us.html

Additionally, after carefully checking every SM title, some other books were recoded as MTB. Finally, books with Law, Code or Dictionary in the title were recoded as reference books (REF). As Vienna lacked a classification of materials, this differentiation was manually performed by two librarians in line with the criteria used by the University of Granada and described above. The error ratio was less than 5%.

For this first pilot analysis, we retrieved the 1000 most borrowed monographs by all types of users since the loan system at each library was implemented (1999–2000 for Vienna, 2001 for Granada). At the University of Vienna and in view of the initial results, it was also possible to collect data on the 1000 monographs most borrowed exclusively by scientists and without considering those from the Faculty of Law library. We then performed analyses and comparisons between all samples at three different levels: book types, language and publisher. Furthermore, for the 100 first most borrowed scientific monographs (SM) and the 100 first most borrowed manuals and textbooks (MTB) citation counts were performed:1) in Web of Science using the "Cited Reference Search", and 2) in Google Scholar using the Publish or Perish software (Harzing, 2007), which calculates various bibliometric indicators based on the data retrieved from this database. All citation data were manually disambiguated and aggregated in both data sources.

Table 1.Loans by document type for the three samples

	TYPE	*TITLES*	*%*	*LOANS*	*%*	*LOANS/TITLE*	*MAX*	*MIN*
Granada	MTB	706	70.6	290943	81.3	412	3922	110
	SM	245	24.5	49465	13.8	202	844	110
	REF	49	4.9	17300	4.8	353	1751	113
	TOTAL	**1000**	**100.0**	**357708**	**100.0**	**358**	**3922**	**110**
Vienna	MTB	334	33.4	166895	27.5	500	2372	207
	SM	200	20.0	60121	9.9	301	1087	206
	REF	466	46.6	380651	62.6	817	7090	207
	TOTAL	**1000**	**100.0**	**607667**	**100.0**	**608**	**7090**	**206**
Vienna Only Scientists	MTB	162	16.2	1690	15.9	10	30	8
	SM	782	78.2	8199	77.2	10	29	8
	REF	56	5.6	738	6.9	13	33	8
	TOTAL	**1000**	**100.0**	**10627**	**100.0**	**10627**	**33**	**8**

Results

A) Loans by document type

The results for both universities are summarized in Table 1. As already mentioned above, the analyses were performed in both universities at edition level and not at title level. In the sample from Vienna (Top 1000, all users) only five books (4 MTB and 1 REF) had the same ISBN and approximately 12% were related to more than one edition (thereof ~ 60% REF). In the second sample (Top 1000, only scientists),only one book (SM/MTB) had the same ISBN twice and only

1.1% of the titles had multiple editions (thereof ~ 60% REF). For the University of Vienna, 119 monographs (~ 10%) comprised both sample 1 and sample 2 (~ 36% SM, ~ 36% REF, ~ 28% MTB). The Pearson correlation between loans to all users and loans to scientists only was rather low (about 0.20).

In the sample from Granada, more than 70% of the most borrowed books (706 titles) were found to be manuals, textbooks and handbooks (MTB), this percentage increasing to 81.3% when taking loans into account. Loans per title for the top MTB were twice the loans per title for top scientific monographs (SM). This category accounts for 24.5% of the loans, while REF books (mainly dictionaries) are just 4.9% of the loans for the Spanish library (Table 1).
The results show a very similar number of scientific monographs in both universities (20% vs. 24%) when considering loans to all users. The differences between the other document types (MTB and REF) can be explained by the large number of books related to law at the University of Vienna (almost 50%), most probably due to their high prices.When considering the number of loans, the rate for scientific monographs drops to 10% in agreement with a lower loan frequency in comparison with the other document types, MTB and REF (see "loans per title" in Table 1). When considering loans to scientists only, the amount of SM grows abruptly to 78%.

Table 2.Loans by language for the three samples

	LANGUAGE	*TITLES*	*% TITLES*	*LOANS*	*% LOANS*	*LOANS/ TITLE*	*MAX*	*MIN*
Granada	Spanish	938	93.8	343935	96.1	367	3922	110
	English	47	4.7	10750	3.0	229	675	133
	Multilingual	7	0.7	1344	0.4	192	431	137
	French	3	0.3	938	0.3	313	649	111
	German	3	0.3	420	0.1	140	157	114
	Italian	2	0.2	321	0.1	161	191	130
	TOTAL	**1000**	**100.0**	**357708**	**100.0**	**358**	**3922**	**110**
Vienna	German	945	94.5	582559	95.9	616	7090	206
	English	55	5.5	25108	4.1	457	1221	206
	TOTAL	**1000**	**100.0**	**607667**	**100.0**	**608**	**7090**	**206**
Vienna Only Scientists	German	653	65.3	7093	66.7	11	33	8
	English	318	31.8	3250	30.6	10	30	8
	French	13	1.3	109	1.0	8	13	8
	Italian	7	0.7	68	0.6	10	10	8
	Serbian	3	0.3	45	0.4	15	20	10
	Spanish	1	0.1	9	0.1	9	9	9
	Lithuanian	1	0.1	9	0.1	9	9	9
	Croatian	1	0.1	17	0.2	17	17	17
	Rumanian	1	0.1	8	0.1	8	8	8
	Czech	1	0.1	10	0.1	10	10	10
	Hungarian	1	0.1	9	0.1	9	9	9
	TOTAL	**1000**	**100.0**	**10627**	**100.0**	**11**	**33**	**8**

B) Loans by language

The results for all three samples are represented in Table 2. Granada and Vienna show very similar results when we consider the language distribution of the top 1000 borrowed books. Almost identical percentage values were reported in both universities: about 94% of the 1000 most borrowed titles are in each country's language, Spanish and German. The percentage of loans for books in national languages is even higher (96%). Despite being the scientific "lingua franca", English is not popular within these academic communities as in both universities the percentage of books in English stays at around 5%. For the only scientists sample in Vienna, this percentage grows to more than 30%, showing an important difference in comparison to the all users sample.

C) Loans by publisher

Tables 3–5 show the top publishers according to the number of titles (or editions) respectively for all three samples.

The alternation of national (Ariel, Pirámide, Manz, Facultas WUV, etc.) and international publishers (McGraw Hill, Prentice, Springer, Pearson, etc.) is similar in both distributions (see Tables 3 and 4). The highest concentration reported for Vienna (only two publishers, Manz and Facultas.WUV account for almost half of the most borrowed books while in Granada ten publishers are needed to surpass 50% of loans) is explained by the predominant role of Manz as the Austrian publisher for law and other reference texts related to law (see also results by document type). In contrast, the names and proportions of the international publishers in both rankings are quite different. For example, McGraw-Hill, top in Granada, does not appear in the top 20 for Vienna, and Springer, top in Vienna, is not present in Granada's top ranking.

When comparing both samples, the publisher distribution is considerably less concentrated for only scientists (~340 publishers) than for all users (~140 publishers for Vienna and ~240 for Granada). Other particularities for the Vienna loans to scientists only (see Table 5) are the appearance of new publisher names, amongst them many foreign university press companies, and the homogeneity of the number of loans per title for all top publishers which fluctuate between nine and 13 loans.

Table 3.Distribution of the top 20 publishers in the Granada sample

PUBLISHER	*TITLES*	*LOANS*	*% LOANS*	*LOANS/TITLE*
McGraw-Hill	105	47690	13.3	454
Ariel	39	12927	3.6	331
Pirámide	37	12736	3.6	344
Prentice Hall	37	17220	4.8	465
Alianza Editorial	35	10511	2.9	300
Edit. Médica Panamericana	32	15701	4.4	491
Síntesis	27	15762	4.4	584
Tecnos	26	14353	4.0	552
Masson	25	6805	1.9	272
Pearson Education	23	10722	3.0	466
Tirant lo Blanch	22	17298	4.8	786
Omega	18	9580	2.7	532
Elsevier	18	7641	2.1	425
Comares	18	4738	1.3	263
Thomson	17	5167	1.4	304
Universidad de Granada	15	4980	1.4	332
Reverté	14	5736	1.6	410
Oxford University Press	13	3991	1.1	307
Addison Wesley	13	6291	1.8	484
Difusión	11	3117	0.9	283
Akal	11	2828	0.8	257
Cátedra	10	1730	0.5	173

Table 4. Distribution of the top 20 publishers for the Vienna sample (all users)

PUBLISHER	*TITLES*	*LOANS*	*% LOANS*	*LOANS/TITLE*
Manz	174	177505	29.2	1020
Facultas.WUV	132	93853	15.4	711
LexisNexis-Verl. ARD Orac	123	66405	10.9	540
Springer	65	47741	7.9	734
Pearson Prentice-Hall	38	19689	3.2	518
Linde	35	17944	3.0	513
Beltz	23	14982	2.5	651
VS Verl. Für Sozialwiss.	19	6847	1.1	360
Thieme	16	6661	1.1	416
Böhlau	16	9520	1.6	595
Hogrefe	16	7996	1.3	500
Oldenbourg	14	8702	1.4	622
SpektrumAkad. Verl.	13	6332	1.0	487
Westdt. Verl.	13	4363	0.7	336
Huber	12	4687	0.8	391
Verl. Österreich	11	7379	1.2	671
Wiley-VCH	11	4336	0.7	394
Leske + Budrich	9	2927	0.5	325
UVK-Verl.-Ges.	9	3386	0.6	376
Elsevier, SpektrumAkad. Verl.	9	3174	0.5	353

Table 5.Top publishers with more than 10 titles for the University of Vienna sample, only scientists

PUBLISHER	*TITLES*	*LOANS*	*% LOANS*	*LOANS/TITLE*
Suhrkamp	46	492	4.6	11
Böhlau	38	423	4.0	11
Oxbow Books	32	295	2.8	9
Manz	28	365	3.4	13
Cambridge Univ. Press	26	253	2.4	10
Campus-Verl.	25	296	2.8	12
Springer	24	302	2.8	13
Oxford Univ. Press	22	214	2.0	10
Beck	21	230	2.2	11
Routledge	20	216	2.0	11
Fink	20	203	1.9	10
Facultas.WUV	18	193	1.8	11
VS Verl. Für Sozialwiss.	15	145	1.4	10
Transcript	13	126	1.2	10
de Gruyter	11	110	1.0	10
Tempus	10	92	0.9	9
Metzler	10	118	1.1	12

Correlations between loans and citations

The mean number of loans for the 100 most borrowed scientific monographs (SM) and the 100 most borrowed manuals (MTB) was 771.6 (Granada) and 640 (Vienna) respectively, showing for both samples a much higher number of loans for those books coded as MTB. Regarding citations, the results show a mean value of 171.5 and 260.7 respectively for Google Scholar, and 25.4 and 53.7 respectively when using Web of Science to retrieve citations. The median of citations was 36 and 18 respectively for the GS sample and only two and 7.5 respectively for the Web of Science sample. When comparing SM and MTB, the differences in loans and citations were statistically significant (CI=95%, $p < 0.05$), MTB median values being much higher than the SM values, with the only exception being Google Scholar citations for the Vienna sample (see Table 6). It is also worth mentioning the sizeable differences between numbers of citations in both databases, with Google Scholar being more exhaustive.

Finally, we performed a correlation analysis by means of the Spearman coefficient (Rho) to test the extent to which loans and citations gathered by both methods were similar. The results show that for the Granada case, there is no correlation at all between loans and citations, regardless of the citation source used and the type of monograph (Table 7). Only correlations between citations gathered from both sources were found to be statistically significant. It is worth noting that this value is higher for scientific monographs (0.765; 0.481 for Vienna) than for handbooks (0.577; 0.466 for Vienna). A statistically significant correlation (0.310) between loans and citations as measured by Google Scholar

was detected for MTB books in the Vienna sample. However, this correlation is so weak that is not appropriate to infer any kind of consistent finding regarding loans and citations. Also, the correlation between citations regarding both databases for MTB and SM books was very weak (less than 0.5).

Table 6. Loans and citations analyses for the 100 most borrowed SM and MTB titles

	MTB	SM	TOTAL
GRANADA			
loans	1245,6 ± 575,7 (1061,5)	297,7 ± 132,4 (254,5)	771,6 ± 632,0 (698,0)
citations_GS	173,9 ± 387,7 (58,5)	169,1 ± 533,1 (25,0)	171,5 ± 465,0 (36,0)
citations_WOS	22,8 ± 66,4 (4,0)	28,1 ± 167,3 (1,0)	25,4 ± 127,0 (2,0)
VIENNA			
loans	908,1 ± 330,8 (841,5)	371,9 ± 110,8 (336,5)	640,0 ± 364,4 (580,0)
citations_GS	244,7 ± 805,1 (7,0)	276,6 ± 1475,5 (23,0)	260,7 ± 1185,7 (18,0)
citations_WOS	62,5 ± 200,2 (12,5)	44,9 ± 138,9 (6,0)	53,7 ± 172,1 (7,5)

Mann-Whitney Test: CI=95%; $p<0.05$. Results are reported as Mean ± Standard Deviation (median)

Table 7. Spearman correlation coefficients for loans, citations and book type

		MTB		SM	
		citations_GS	citations_WOS	citations_GS	citations_WOS
GRANADA					
loans	Rho	0.135	0.081	0.099	0.115
	Sig.	0.181	0.421	0.328	0.254
citations_GS	Rho		0.577*		0.765*
	Sig.		0.000		0.000
VIENNA					
loans	Rho	0.310*	0.189	0.032	0.060
	Sig.	0.002	0.059	0.751	0.550
citations_GS	Rho		0.466*		0.481*
	Sig.		0.000		0.000

*Correlation is statistically significant at 0.01 level (2-tailed)

Discussion and concluding remarks

In this paper we analyse whether library loans might be used as a proxy for the measurement of monograph use and the feasibility of using library loans as a possible selection criterion for including monographs in citation indexes. To this end, we conducted an exploratory study analysing loan data and citation data from two university libraries which represent two non-Anglo-Saxon academic communities, a Spanish-speaking community and a German-speaking community.

Methodologically, this has not been an easy process. A number of technical factors need to be considered before taking loans as a valid measure for the useof monographs and subsequently as a selection criterion for book citation indexes:

the publication year is not the acquisition year, extensions, loan times or loan counts, differentiation of the user types and materials, different counts in different libraries, multiple editions and copies, etc. Also, the loan time can differ between universities and types of users and materials. Finally, some books could be classified as not for loan due to several reasons, so no data could be gathered for them. It is also worth mentioning the presence of library departments which loan books that may not be within the general automated library system.

As shown in our study, different approaches can be taken, such as measuring numbers of copies, editions or titles of monographs. Also, the aggregation of counts from different editions and translations should be considered, as a number of the most "popular" books happen to be translations of Anglo-Saxon monographs. Counting only the translations could miss the academic impact of a book as a whole. One additional technical difficulty is the detection of citations referring to different books where the title coincides in various languages, as happens for titles such as Economia (Economy) or Biologia (Biology) in Spanish, Portuguese and Italian. Usually handbooks and manuals have a broad coverage, so the titles are very short (one or two words in many cases), which makes it impossible to split citations for every language.

Most of the top books were manuals, handbooks and textbooks, or reference books such as those for law and dictionaries.This is understandable as the main users of university libraries are students. The results also show that scientific monographs in both universities account for 20%–25% of the most borrowed books when considering loans for all users. However, this percentage increases to 78% when analysing loan patterns for scientists only. This fact points to the need to differentiate between types of users to assess more precisely the reliability of loans as a usage indicator for monographs.

A second conclusion is that both academic communities (Vienna and Granada) prefer to borrow books in their respective languages, regardless of the original language of the publication. The outstanding percentage (around 95% for Spanish and German) of loans for publications in national languages could be explained by the type of users, mainly students. When assessing the loan behaviour of scientists only (solely for the Vienna sample), we have found that they are more likely to borrow English monographs than the other user groups. This is not surprising as these monographs are expected to convey more specialized information which could justify the lack of a translation of such books. However, for the Vienna scientists, German is also the preferred language when looking up information in scientific monographs, with more than 65% of the most borrowed books being in German.

The rankings of publishers according to the most borrowed books also show the presence of both international publishers and local well-known publishing houses.

Both of these tend to distribute scientific monographs and handbooks with a broad orientation and these are the materials most borrowed from academic libraries.

We also have found important differences in the number of citations retrieved using Web of Science and Google Scholar. Google Scholar has gathered more citations than Web of Science for most of the samples when considering the median number of citations. The exception is the Vienna sample for MTB where Web of Science retrieves a higher number of mentions with respect to those monographs. Standard deviations for citation statistics suggest that within the most borrowed books, highly cited books coexist with other monographs that are not noticed by the scientific community.

Regarding citations relating to the most borrowed books, it is important to mention the lack of correlation between these two variables in our study for the Granada sample and the very weak correlation for Vienna MTB books when using Google Scholar. The fact that the books cover a broad range of topics, mainly in languages other than English and of a general nature, may be important for the interpretation of this finding. We also have to consider the different citation behaviours in each discipline and the aging of books, further issues which may affect these results, along with the aforementioned technical difficulties. These considerations also lead us to think that a discipline-focused study could shed more light on the validity of loans as a criterion for selecting monographs in selective indexes than could a broad study.

This study also confirms the need for facilities to aggregate different editions and translations of a certain book under one record in order to indicate the actual relevance of the overall work. It could be a future task for academic libraries to provide usage data, especially concerning book loans for regular systematic analyses, as they still constitute an important aspect of the scholarly communication process, though rarely recognized by bibliometric studies to date.

Acknowledgments

We would like to thank the Library of the University of Granada for facilitating the provision of the data for the purposes of this study. Nicolás Robinson-García is currently supported by an FPU grant from the Ministerio de Economía y Competitividad of the Spanish Government.

References

Cronin, B., Snyder, H. & Atkins, H. (1997). Comparative citation rankings of authors in monographic and journal literature: a study of sociology. *Journal of Documentation*, 53(3), 263–273.

Giménez-Toledo, E., Tejada-Artigas, C. & Mañana-Rodríguez, J. (2013). Evaluation of scientific books' publishers in social sciences and humanities: results of a survey. *Research Evaluation*, 22(1), 64–77.
Gorraiz, J., Purnell, P. & Glänzel, W. (2013).Opportunities and limitations of the Book Citation Index. *Journal of the American Society of Information Science and Technology*, in press
Gorraiz, J. & Schlögl, C. (2006). Document delivery as a source for bibliometric analyses: the case of Subito, *Journal of Information Science*, 32(3), 223–237.
Harzing, A.W. (2007).*Publish or Perish*. Available at:http://www.harzing.com/pop.htm
Hicks, D. (1999).The difficulty of achieving full coverage of international social science literature and the bibliometric consequences. *Scientometrics*, 44(2), 193–215.
Hicks, D. (2004). The four literatures of social science.In Moed, H.F., Glänzel, W. &Schmoch, U. (Eds.). *Handbook of Quantitative Science and Technology Research: The Use of Publication and Patent Statistics in Studies of S&T systems*. Kluwer Academic Publishers, Netherlands, pp. 473–496.
Leydesdorff, L. & Felt, U. (2012). Edited volumes, monographs and book chapters in the Book Citation Index (BKCI) and Science Citation Index (SCI, SoSCI, A&HCI). *Journal of Scientometric Research*,1(1), 28–34.
Linmans, A.J.M. (2010). Why with bibliometrics the humanities does not need to be the weakest link. Indicators for research evaluation based on citations, library bindings and productivity measures. *Scientometrics*, 83(2), 337–354.
Nederhof, A.J. (2006). Bibliometric monitoring of research performance in the social sciences and the humanities: areview. *Scientometrics*, 66(1), 81–100.
Price, D. de S. (1963). *Little Science, Big Science*. New York, Columbia University Press.
Priem, J., Taraborelli, D., Groth, P.&Neylon, C. (2010).*Almetrics: a manifesto*. Available at: http://altmetrics.org/manifesto/
Research Information Network (2009). *Communication knowledge: how and why UK researchers publish and disseminate their findings*. Joint Information Systems Committee.Available at: http://www.rin.ac.uk/communicating-knowledge
Research Information Network (2011).*Reinventing research? Information practices in the humanities*. Joint Information Systems Committee.Available at: http://www.rin.ac.uk/system/files/attachments/Humanities_Case_Studies_for_screen_2_0.pdf
Torres-Salinas, D.&Moed, H.F. (2009).Library Catalog Analysis as a tool in studies of social sciences and humanities: an exploratory study of published book titles in Economics. *Journal of Informetrics, 3*(1), 9–26.
Torres-Salinas, D., Robinson-García, N., Jiménez-Contreras, E. & Delgado López-Cózar, E. (2012).Towards a 'Book Publishers Citation Reports'. First

approach using the 'Book Citation Index'.*Revista Española de Documentación Científica*, 35(4), 615–620.
Torres-Salinas, D., Rodriguez-Sánchez, R., Robinson-García, N., Fdez-Valdivia, J. & García, J.A. (2013).Mapping citation patterns of book chapters using the Book Citation Index. *Journal of Informetrics*, in press.
Testa, J. (2011) *The book selection process for the Book Citation Index in Web of Science*. Available at: http://wokinfo.com/media/pdf/BKCI-SelectionEssay_web.pdf
University of Granada (2011).*Memoria básica de investigación 2011*. Granada.
University of Granada (nd). *Anuario de la Biblioteca de la UGR. Año 2011*. Granada.
White, H., Boell, S.K., Yu, H., Davis, M., Wilson, C.S. & Cole, F.T.H. (2009). Libcitations: A measure for comparative assessment of book publications in the humanities and social sciences. *Journal of the American Society of Information Science and Technology*, 60(6), 1083–1096.
Zuccala, A., & van Leeuwen, T. (2011). Book reviews in humanities research evaluations. *Journal of the American Society of Information Science and Technology*, 62(10), 1979–1991.

MOTIVATION FOR HYPERLINK CREATION USING INTER-PAGE RELATIONSHIPS

Patrick Kenekayoro[1] Kevan Buckley[2] Mike Thelwall[3]

[1] *Patrick.Kenekayoro@wlv.ac.uk* [2] *K.A.Buckley@wlv.ac.uk* [3] *M.Thelwall@wlv.ac.uk*
Statistical Cybermetrics Research Group, University of Wolverhampton, Wulfruna St Wolverhampton, West Midlands WV1 1LY (UK)

Abstract
Using raw hyperlink counts for webometrics research has been shown to be unreliable and researchers have looked for alternatives. One alternative is classifying hyperlinks in a website based on the motivation behind the hyperlink creation. The method used for this type of classification involves manually visiting a webpage and then classifying individual links on the webpage. This is time consuming, making it infeasible for large scale studies. This paper speeds up the classification of hyperlinks in UK academic websites by using a machine learning technique, decision tree induction, to group web pages found in UK academic websites into one of eight categories and then infer the motivation for the creation of a hyperlink in a webpage based on the linking pattern of the category the webpage belongs to.

Keywords:
webometrics, decision tree induction, link classification, supervised learning

Introduction
Webometrics has been defined as "*the study of web based content with primarily quantitative research methods for social science goals using techniques that are not specific to one field of study"* (Thelwall, 2009). Techniques from different fields like mathematics and statistics have been applied to study web content for webometrics research. Machine learning is an area in computer science that is concerned with pattern discovery. This technique has not been used extensively in webometrics research, but has been applied in several computing web studies, for example (Chau & Chen, 2008; Luo, Lin, Xiong, Zhao & Shi, 2009; Qi & Davison, 2009).
A particular area where machine learning can be applied to webometrics research is in link analysis. Link analysis involves the study of link relationships between a group of websites or the link structure of a group of websites. It has been successfully used as a source of business intelligence (Vaughan & Wu, 2004; Vaughan, 2005) and used in several studies of academic websites (Thelwall, 2002c; Thelwall & Wilkinson, 2003). Nevertheless, using raw link counts between websites can be unreliable and several researchers have attempted to find alternatives to raw link counting or tried to understand the meaning of link counts between websites. Researchers have classified links in the web pages of academic

institutions as research or non-research related (Thelwall, 2001), substantive or non-substantive (Smith, 2003) and shallow or deep (Vaseleiadou & van den Besselaar, 2006). Other researchers have tried to identify the reasons why links in academic web pages were created (Bar-Ilan, 2004; Wilkinson, Harries, Thelwall, & Price, 2003) but there is no agreement about an effective way to classify the reasons for university interlinking on a large scale, which is the goal of this paper.
The manual classification of individual links is time consuming, thus it is infeasible for large scale studies. Perhaps this is the reason why there is a dearth of literature in link classification. In this paper, the reason for link creation is inferred from the relationship between the two pages the hyperlink connects; the source page and the target page. Hyperlinks have been previously classified using the target page (Thelwall, 2001), using both the source and target page can give better insight to the reason for hyperlink creation and web page classification is a simpler problem than individual link classification.
Typically, a university's website has thousands of web pages so an effective approach will be to group similar web pages together and then infer why a link is created based on the link creation motivation of the group that web page belongs to. This makes it necessary to identify the types of web pages that can be found in a university's website. Page types are identified using the mission of a university and the function of its website as a guide.
Stuart, Thelwall and Harries (2007) suggest that methods for automatic classification of hyperlinks should be developed if links are to be fully harnessed for webometrics research. The reason why a link is created, that is the relationship between the source and target page can form classes of links in a university's website. Machine learning techniques are used to automate the classification scheme, there by bringing us a step closer to fully harnessing hyperlinks for webometrics research.
The main goal of this study is to identify a method that effectively determines the reason why a link in a UK university's website has been created. This is achieved by grouping web pages into categories that are in line with the three missions of Higher Education Institutions (HEIs), automating the classification scheme with machine learning techniques and then examining the relationship between page categories. This paper begins with background information on link classification, the supervised learning technique used in this study, then, categories of pages that could be found in UK university's websites are identified, and the results of classification with a supervised learning technique are shown and the relationship between a random sample of web pages analysed.

Link Classification

Webometrics can be used as a source of business intelligence. Vaughan and Wu (2004) showed that a link count to a company's website positively correlates with the company's business performance and co-linked web pages were used to identify a company's competitors (Vaughan, 2005). Webometrics has also been applied in the study of academic institutions (Payne & Thelwall, 2004; Thelwall

& Wilkinson, 2004; Vaughan & Thelwall, 2005) and in the identification of political trends (Park & Thelwall, 2008; Romero-Frías & Vaughan, 2010; Romero-Frías & Vaughan, 2012).

Raw link counts are unreliable (Thelwall, 2002a; Thelwall, 2002b) because links are prone to spamming (Smith, 1999) and there could be different motivations behind the creation of hyperlinks (Wilkinson et al., 2003). There are different reasons behind link creation, it can vary according to the function and operational relationship between the different organisations studied (Minguillo & Thelwall, 2011). Linking between municipalities in Finland is motivated by cooperation made possible because of geographic closeness (Holmberg & Thelwall, 2009) and co-linking of business web pages tends to be for business reasons (Vaughan, Kipp, & Gao, 2007) which is different from co linking of university web pages that could just be as a result of general reference. This difference highlights the disparity between linking patterns in different domains, thus caution should be used when applying a method tested in one domain to a different domain.

Several attempts have been made to classify web pages in a University's website but there is no consensus as to how links can be classified and Thelwall (2006) suggests no single link interpretation is perfect. Two approaches to hyperlink interpretation are (Thelwall, 2006)

- Interviewing a random selection of link creators about why they created a link
- Classification of a random selection of links in a way that is helpful to the research goals

Author interviewing may give a more accurate result but classification of links is a more practical approach; it is the most common method for hyperlink classification in the academic web space (Bar-Ilan, 2004; Bar-Ilan, 2005; Thelwall, 2001; Thelwall, 2003; Wilkinson et al., 2003).

Thelwall (2001) classified hyperlinks in web pages of UK academic institutions as research related or not research related based on the content of the target page, which he noted was a practical step. Although classification of some pages was subjective, a situation similar for all research in this area, some general rules were created to for the classification process. For example, departments' homepages, staff research profiles, web pages of research groups were classified as research related while electronic journal pages were classified as non-research related. Results showed that using only research related links increased the correlation with average research rating of UK institutions.

Wilkinson and his colleagues (2003) studied 414 random links between UK academic institutions in order to identify the motivations for hyperlink creation. Even though individual links were investigated, the reason from link creation was determined using the source page and target page. They suggest that this approach is difficult as it is impossible to guess the motivation for link creation and in some cases there could be several motivations.

Thelwall (2003) studied 100 random inter site links from a UK university's' web page to the homepage of another UK university. He grouped web pages into four categories: *navigational*: a link created to direct uses to other non-subject specific information, *ownership*: links to partners and they were often in the form of a clickable image of the university's crest, *social*: links to institutions of collaborating research groups and *gratuitous*: links created without any specific motivation.
Bar-Ilan (2004), in perhaps the most systematic study so far, classified the link, source page and target page from different aspects, link context, link tone and several other properties, in a case study of eight universities in Israel. This approach is difficult as Wilkinson and his colleagues (2003) pointed out problems with guessing author motivations and subjective decisions in the classification process. It is also impractical to study each link individually because of the sheer number of links that could be found in a university's website. A more practical approach to link classification is finding the relationship between the two pages a hyperlink connects, the source and target page, which reduces the problem of hyperlink classification to web page classification. To study the relationship between web pages, the type of pages that could be found in a university's website must be identified.

Supervised Learning

Although web page classification is simpler than individual link classification, this process is still infeasible for large scale manual studies because a typical university's website can have thousands of web pages.
Supervised machine learning is a Computer Science technique concerned with teaching machines to predict unseen cases of input data based on patterns identified from previously observed examples, usually called a training set. There are several machine learning algorithms, like decision tree induction, support vector machines, neural networks and k nearest neighbours' classifiers. Decision tree induction has an advantage over other models in that it is easy for humans to understand the resulting classifier, because of its high level rules, as opposed to other black box models like neural networks whose model is encapsulated in a complex numerical model. For this reason, decision tree induction was used in this study, but other classification techniques may produces similar or better results.

Decision tree induction

Decision tree induction recursively splits data into disjoint sets according to a criterion. Each node is a feature an instance can have, and leaf nodes contain output classes for instances that reach that node. Figure 1 is an example of a decision tree classifier that classifies instances into one of three possible classes. Classification of instances start at the root node, then the instances traverse down the tree in the direction that meets several criteria until a leaf node is reached. The value of the leaf node is then assigned to that instance.

Constructing optimal binary decision trees is a NP-complete problem (Kotsiantis, Zaharakis, & Pintelas, 2006), however several techniques like the C4.5 algorithm (Quinlan, 1993) and CART, acronym for Classification And Regression Trees (Breiman, Friedman, Stone, & Olshen, 1984) can be used to build decision trees. CART and C4.5 algorithms are implemented in a machine learning toolkit, WEKA (Hall et al., 2009) that is used in this study to automate the classification scheme.

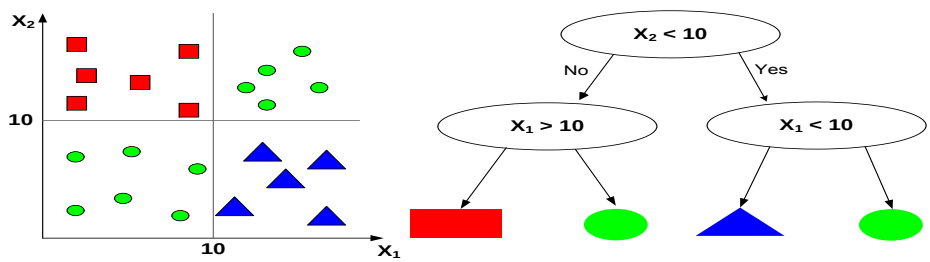


Figure 1 A Decision Tree Classifier

Two major phases of a decision tree induction are the growth phase and the pruning phase (Kotsiantis, 2011). The growth phase involves splitting the training data into disjoint sets and the pruning phase reduces the size of the decision tree to avoid overfitting.

Decision tree induction Pseudo Code (Kotsiantis, 2007)

```
1. Check for base cases
2. For each attribute a
3.    Find the feature that best divides the training
   data
4. Let a_best be the attribute that best splits data
5. Create a decision node that splits on a_best
6. Recurse on the sub-lists obtained by splitting on
   a_best and add nodes as children
```

A major difference between the C4.5 and CART algorithms is the way the best feature that separates the training data is selected. The attribute with maximum information gain is used to split the training set. C4.5 uses entropy to compute the information gain, while CART uses the Gini index. The entropy is calculated by:

$$Entropy(S) = -\sum_{i=1}^{n} Freq(C_i, S) * log\big(Freq(C_i, S)\big)$$

$Freq(C_i, S)$ is the relative frequency of instances in class C_i.

The Gini index is computed by:

$$GiniIndex(S) = 1 - \sum_{i=1}^{n} Freq(C_i, S)^2$$

And information gain is computed by:

$$InformationGain(S, A) = I(S) - \sum_{i=1}^{n} \frac{|S_i|}{|S|} * I(S_i)$$

The formula above computes the information gain of attribute A in data set S where $i = 1 \ldots n$ are possible values of attribute A, $S_i \ldots S_n$ are partitioned subsets of S where attribute A is$i$, $I(S)$ is the entropy in C4.5 algorithm and Gini index in CART algorithm.

Testing and Evaluation

It is essential to evaluate the accuracy of a learning algorithm. The fundamental assumption of machine learning is that the distribution of training data is identical to the distribution of test data and future examples (Liu, 2006). If the learning algorithm generalizes the training set, then the machine learning assumption suggests that it will perform well for future unseen examples. Generalization is estimated by the accuracy of the learning algorithm, measured by the equation:

$$Accuracy = \frac{Number\ of\ correctly\ classifed\ examples}{Total\ number\ of\ examples}$$

Ultimately, the accuracy measure depends on the application which applies the learning model. Precision, recall and F-measure give a more elaborate description of the performance of a learning algorithm. They are calculated based on four parameters: True Positives (TP), False Negative (FN), False Positive (FP) and True Negative (TN). Given the test set D, if each instance of the set can be of class $y = (1, -1)$ and $f(x)$ is the function trained to predict future unseen instances. The parameters TP, FN, FP and TN are:

- True Positives (TP) = $Count\ (f(x) = 1\ and\ y = 1)$
- False Negatives (FN) = $Count\ (f(x) = -1\ and\ y = 1)$
- False Positives (FP) = $Count\ (f(x) = 1\ and\ y = -1)$
- True Negatives (TN) = $Count\ (f(x) = -1\ and\ y = -1)$

Precision, recall and F-measure is computed by:

$$Precision = \frac{TP}{TP + FP}\ ;\quad Recall = \frac{TP}{TP + FN}\ ;\ F_{measure} = \frac{2 * Precision * Recall}{Precision + Recall}$$

Precision and recall are used when the interest is on one particular class. In most cases, the accuracy formula earlier defined as percentage of correctly classified divided by the total number of instances is used.

A rule of thumb in machine learning is that the input data should be divided into two-thirds for training and the remaining one-third for tests/validation. Another technique, cross validation is also used. In cross validation, the input data is divided into n equal disjoint subsets. Each subset is used as the test set, and the union of the others the training set. The accuracy of the classifier is the average accuracy of the n different subsets. A special case of cross validation is the leave one out approach, where a single example is used as test and all others for training. In this case, n is the number of training examples, thus this method can be computationally expensive.

Page Types

Bar-Ilan (2005) created a detailed framework for characterising the uses of links in an academic website. The scheme had 31 possible relationships between two pages and 13 different page types. Using machine learning to automate this classification scheme can be difficult because of the number of variables involved. Moreover, this study aims to group web pages based on the three missions of Higher Education Institutions (HEI) and the functions of its website. Bar-Ilan's (2005) method does not fit into this classification scheme. For example, a physical unit can comprise of an administrative unit or a research unit. These two units serve different purposes in regard to the missions of HEIs so they should not be grouped together. Because of this, this study uses a classification scheme that is less detailed than (Bar-Ilan, 2005), but is easier to automate with machine learning, and is more in line with the aims of this paper.

If the links in a university's websites are to be classified based on the relationship between two pages, it is necessary to identify the types of pages that can be found in a university's website, and then study how these page types interlink.

In order to identify the type of pages in a university's website a random set of web pages in a university's domain were visited and manually classified. A custom web crawler was designed to get the link structure of 111 UK universities. The crawler extracted links originating from a UK university to another UK university, not visiting all pages in a university's website. It only covers links that can be reached by following links from a university's homepage, similar to Thelwall (2001). An additional constraint was added as this work is only concerned with hyperlinks between UK academic institutions. New web pages were not added to the list of websites to visit when the crawler visited 2000 consecutive pages without finding a link to another UK university. 15 link pairs between universities were randomly selected and then the web pages that these links direct to were used to identify the page types in a university's website.

Websites of organisations are designed to disseminate the activities and functions of that organisation. In some cases the structure of the website replicates the physical structure of that organisation. Nowadays, Higher Education Institutions (HEIs) have three main goals, teaching, research and what authors simply refer to as the third mission. If websites of universities are designed to channel the activities and functions of that university, which in turn are in-line with the (three)

main goals of HEIs then university websites will be similar. From a university's homepage a general idea about the goals of the university as well as the function of its website can be inferred. Thus text from the homepages of UK universities was used to determine the types of pages studied in this work. Text from the homepages of UK universities was extracted, and then the top 20 words when stop words were removed were used as a guide to determine preliminary page types.

Table 1 Page Type and description

Page Type	Description
About	Promotes the school and gives information to staff/students. Examples of such pages are news, information, university profile, prospectus, events
Business and innovation	Connects the school to non-academic environment. Examples include expert services offered, community projects, partnership with science parks
Discussion	Forums, blogs or web page containing opinions of a user. Comments or posts in these pages are for a variety of reasons; research, teaching or recreational.
Support	Contains a repository of learning resources for students/staff support, skills for learning, services, counselling. Examples include Archives, Books, Database.
Research	Involved with the production of new knowledge. Examples include Research centres, research groups, research projects, academic schools/departments, conferences, Abstract, Academic Article
Staff	Related to a staff in the university. Examples include staff homepage, staff profiles, list of publication, CV
Student Life	Enhances the student experience. Examples include student union website, student benefits, campus facilities, tourism, recreation
Study	Involved with transfer of knowledge. Examples include module learning materials, module timetables, module page, lectures

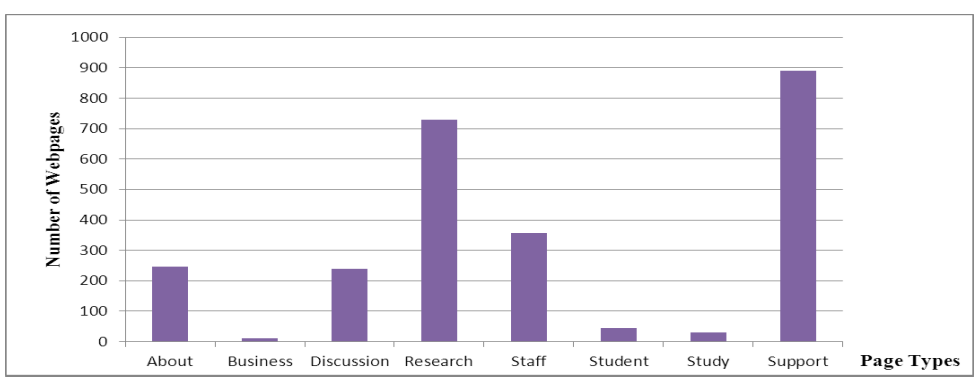


Figure 2 Distribution of page types in 2500 random UK university web pages

100 random web pages were given to an independent researcher for classification. When a page type identified by the independent researcher did not fit into the preliminary page types, a new category was created.
The page types in Table 1 are largely based on the authors' opinion; top words were only used to assist in the decision process. These page types however, cover the majority of web pages found in university's websites. Pages were grouped into one of the eight categories and then the possible reasons for link creation were identified for links originating from one page type to another.

Automatic Classification

Decision trees are constructed using features of the training set. Different features may associate a training instance to a particular page type. In this case, training instances are the web pages to be classified. 2500 web pages were randomly selected and manually classified into one of eight categories a UK university's web page could belong to. Two thirds of the web pages were used for training and the rest for testing.
The features of each web page were derived from the web page title and/or web page URL pre-processed and then represented as a word vector of TF (term frequencies) or inverse document frequency multiplied by the term frequency (TFIDF). In this case, a term frequency representation is similar to a binary representation because page titles are short thus words rarely occur more than once.
Pre-processing transforms the text to an information rich representation. Pre-processing steps used are:
Word Tokenization: Splits attributes (web page title/URL) into word tokens. For example "the quick brown fox jumps over the lazy dog" has 9 (nine) tokens of 8 (eight) word types. A simple way to achieve tokenization is by assuming [space] separates word tokens. Only words that did not contain any non-alphabetic characters were used in this study.
Capitalization: All characters were represented in lower case.
Stop word removal: Removal of the most frequent words that occur in the English language. Words like "the, and, a, is ... " are all removed. WEKA (Hall et al., 2009) contains the list of stop words that was used in this study. The 111 university names as well as www, http and https were added to the list of stop words.
Stemming: Stemming reduces inflected words to their root form or stem. For example jumps and jumping have the same stem, jump. Accuracy may be improved if all words are represented in their root form. The Porter Stemmer is a commonly used stemming algorithm and it is used in this study. Stemming algorithms occasionally make errors. For example, the Porter Stemmer stems both university and universe to univers.
WEKA (Hall et al., 2009) implements decision tree induction in its J48 algorithm. Although the default settings may give satisfactory results, in some cases tweaking the settings may improve the accuracy of the algorithm. Feature

selection as well as data pre-processing is also an important aspect in supervised learning. Table 2 shows how pre-processing options influence the accuracy of the classifier. Training accuracy is determined using a 10 fold cross validation, while verification is determined using the formula $Verification = \frac{Number\ of\ correctly\ classifed\ examples}{Total\ number\ of\ examples}$; the data set used in verification is different from the set used in training. Verification gives an estimate of the out of sample performance of the learning algorithm and is also used to identify overfitting. A machine learning algorithm overfits when it performs better in training than in testing.

Table 2 Training and Verification of top 10 pre-processing options of decision tree induction

Bigrams/ Unigrams	TF *IDF	Stem	Stop words	Page title	URL	Training	Verification
Unigrams		Yes	Yes	Yes	Yes	72.13	71.25
Unigrams	Yes	Yes	Yes	Yes	Yes	72.16	71.25
Unigrams	Yes	Yes		Yes	Yes	71.15	69.9
Unigrams		Yes		Yes	Yes	71.16	69.9
Bigrams + Unigrams			Yes	Yes	Yes	71.57	69.66
Bigrams + Unigrams	Yes	Yes	Yes	Yes	Yes	72.31	69.66
Bigrams + Unigrams		Yes	Yes	Yes	Yes	72.32	69.65
Unigrams				Yes	Yes	71.86	68.92
Unigrams	Yes			Yes	Yes	72.66	68.92
Unigrams			Yes	Yes	Yes	72.78	68.55

Settings of the classifier were tweaked and the best results are shown in Table 2. On average, the top 250 features were used in the construction of the decision tree.

When word counts are used as features, the number of features increases as the training set increases. Too many irrelevant features affect the speed as well as accuracy of a learning algorithm, so input features have to be carefully selected. The J48 algorithm uses the information gain to select the best feature that optimally splits the training data, so it is logical to use information gain to identify relevant features. Other methods like principal component analysis and entropy-weighted genetic algorithm can also be used to reduce the size of the features. Another way to reduce the feature size is by generation an initial decision tree using all features, and then excluding those features not used in the initial tree during subsequent training. In tests, only 7 percent of the features were used in the final decision tree. Excluding features that were not used produced a slight improvement in the accuracy of the decision tree.

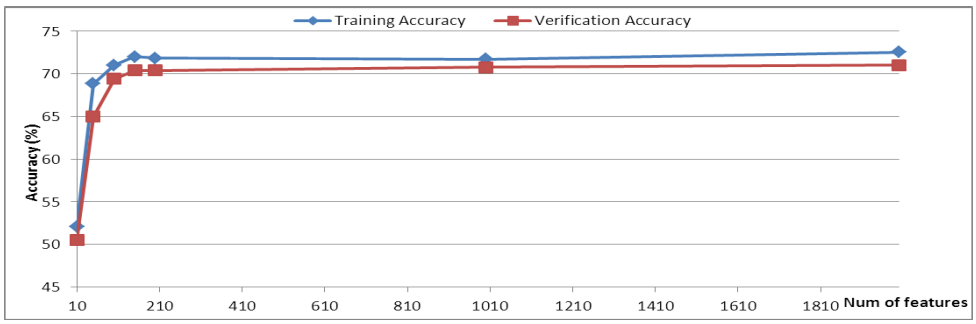


Figure 3 Influence of feature size on accuracy of the classifier

Figure 3 shows how the number of features affects the accuracy of the decision tree classifier when the best pre-processing setting from Table 2 was tested for different number of input features. Initially, as the feature size increases, the accuracy of the classifier also increases but at some point the increase in size doesn't improve the accuracy; it only reduces the speed of the decision tree classifier.

Table 3 Classification Accuracy of each page type

Class	Precision	Recall	F Measure
About	0.59	0.46	0.52
Business and Innovation	0.00	0.00	0.00
Discussion	0.87	0.89	0.88
Research	0.63	0.80	0.70
Staff	0.78	0.75	0.77
Student Life	0.63	0.29	0.40
Study	1.00	0.33	0.50
Support	0.78	0.71	0.74

The overall accuracy of the decision tree classifier is about 71%. However it is interesting to know how accurately the classifier identifies individual page types. This is determined using precision, recall and F measure in Table 3. In summary, precision is the likelihood that the classifier will correctly classify a web page of type X as class X, while recall is the likelihood that the classifier will not classify a web page that is not of type X as class X. F measure is the accuracy of an individual class computed by a formula that depends on the precision and recall.
Figure 4 shows a partial decision tree for the classification of the web pages. This is not the optimal result that can be achieved. The settings of the classifier were adjusted to reduce the size of the tree. If the decision tree in Figure 4 is used to classify university web pages, several page types will always be incorrectly classified. Business and Innovation, Study and Student Life pages do not appear in any leaf node so they will always be misclassified. Web pages that do not

contain any of the keywords will be classified as research pages. The tree in Figure 4 has an accuracy of 46.8%.
The nodes in the tree above show top terms in the feature set that are associated with a specific page type.

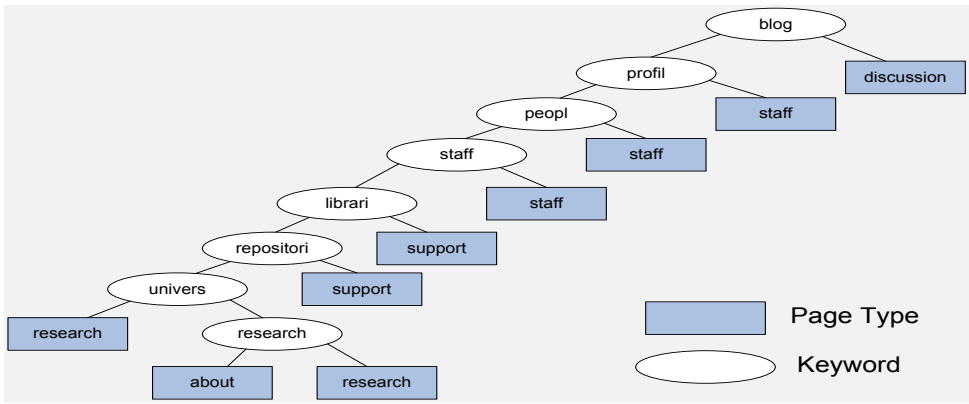


Figure 4 Tree view of a decision tree induction classifier

Inter-page Relationships

Each page type was studied to identify the type of pages they link to and possible reasons for interlinking. 15 random links were randomly selected but at the time of the investigation, some of these pages, either the source or target pages were not available online. Such links were excluded from the study.
With eight page categories, if each category has a link to all other categories including itself there becomes $8 * 8 = 64$ page relationships to study. However, not all page types interlink and majority of links are between the same page types. Subsequent sections describe results for each page type.

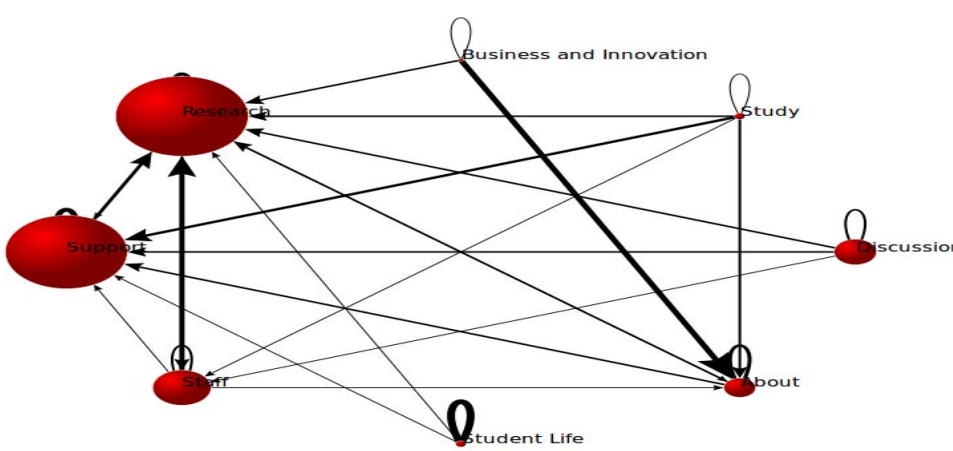


Figure 5 Visual representation of interlinking between page types

Table 4 Page types and reasons for link creation

Page Type	Size	Additional Notes
Support	35%	❖ Rarely link to other page types ❖ Links to research pages that own or created the resource in the support page. ❖ Links are created to direct users to other relevant information, often to other pages that are created to improve learning, research or teaching skills.
Research	28.6%	❖ Links to About pages, usually a clickable logo of a collaborating university; organisational links as described by Thelwall (2003). ❖ Pages about research projects had links to staff pages of its collaborators or homepages of research groups or department. ❖ Research pages had links to all research groups or departments in the same scientific field.
Staff	14%	❖ Links to about pages (homepages of universities) were often what Thelwall (2003) refers to a gratuitous links. ❖ Links to support pages that contain a resource, for example, staffs' publications. ❖ Links to other staff pages because of collaboration in a research project or co-authorship in a publication.
About	9.7%	❖ Are linked to but rarely link to other universities. ❖ Largely made up of course prospectus and university homepages. ❖ Majority of outgoing links are for non-scholarly reasons.
Discussion	9.4%	❖ Links are created for a variety of reasons, so it is very difficult to identify a general pattern. ❖ Each blog entry belongs to a particular page type, and reasons for linking are the same as reasons of its corresponding page type.
Student Life (SL)	1.8%	❖ Mainly Link to SL pages in close geographic locations. ❖ A part of the reason why link analysis research shows that UK universities links to other universities in close geographic location.
Study	1.2%	❖ Majority of links are to support pages containing information relevant to the course. ❖ Links to research pages that contained software/ research output used in the course. ❖ Links to staff pages that authored course material, or a visiting professor ❖ Represents only a small set of total pages, perhaps because teaching materials are located on a protected server, thus inaccessible through public web crawlers.

Figure 5 shows how different page types inter links. Vertices represent page types and arcs represent links from a page type to another page type. The size of the vertices indicates the number of web pages in that page type, while the colour of arcs indicates the percentage of link from a page type to another page type.

Thicker arcs mean a large percentage of links from that page types go to the page type on the other end of the arc, while bigger vertices mean a large percentage of web pages belong to this page type.
Random links are manually classified in order to identify linking patterns of different page types. Table 4 shows linking patterns identified for different page types.

Conclusion and Further Work

Hyperlink classification based on inter-page relationships is a practical solution compared to other methods that try to classify individual links on web pages. This work used the source and target page to determine the reason for linking. The relationship between the source and target pages was similar for web pages in the same category.
Support pages had links to its resource creator or pages that gave additional information that enhances teaching or learning schools. Staff web pages about a research project are ideal to identify collaboration or cooperation between institutions. Other research pages only show the amount of research activity going on in the university. Web pages about the living experience in the university, even if they represent only a small part link to web pages in close geographical regions. These links are for non-scholarly reasons and should be excluded when identifying academic relationships between universities. Other page types, business and innovation and study web pages contributes less than 1 percent to the total links to other universities, perhaps because they are situated in an area not accessible by web crawlers or link to other non-academic originations. Administrative pages also contained few links to other universities, and they were for non-scholarly reasons. However, majority of links to administrative pages were either gratuitous or as a result of collaboration.
Even though classification based on inter page relationship is more practical than classifying individual links, it is still infeasible to manually classify each page. Web pages are fewer than hyperlinks, but they cannot be efficiently classified manually because a typical UK university website contains thousands of web pages. Classification of page types can be automated using a supervised machine learning technique; decision tree induction was used in this study and results showed moderate accuracy; 71%.
Links between two staff pages suggest collaboration, and as supervised learning methods can automatically identify staff pages with decent accuracy; F measure of 77 percent, there is a possibility for more in depth analysis of inter linking between universities staffs so this type of links can be in cooperated to bibliometric analysis.
There are several other machine learning algorithms which may give more accurate results. Further work will aim to compare results of other classification techniques as well as apply natural language processing techniques to improve the accuracy of the classifier.

References

Bar-Ilan, J. (2004). A microscopic link analysis of academic institutions within a country ” the case of israel. *Scientometrics, 59*(3), 391-403. Retrieved from http://dx.doi.org/10.1023/B:SCIE.0000018540.33706.c1

Bar-Ilan, J. (2005). What do we know about links and linking? A framework for studying links in academic environments. *Information Processing & Management, 41*(4), 973-986. doi:10.1016/j.ipm.2004.02.005

Breiman, L., Friedman, J., Stone, C. J., & Olshen, R. A. (1984). *Classification and regression trees*
Wadsforth International Group.

Chau, M., & Chen, H. (2008). A machine learning approach to web page filtering using content and structure analysis. *Decision Support Systems, 44*(2), 482-494. doi:10.1016/j.dss.2007.06.002

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., & Witten, I. H. (2009). The WEKA data mining software: An update. *SIGKDD Explor.Newsl., 11*(1), 10-18. doi:10.1145/1656274.1656278

Holmberg, K., & Thelwall, M. (2009). Local government web sites in finland: A geographic and webometric analysis *Scientometrics, 79*(1), 157 - 169. doi:10.1007/s11192-009-0410-6

Kotsiantis, S. B. (2007). Supervised machine learning: A review of classification techniques. Paper presented at the *Proceedings of the 2007 Conference on Emerging Artificial Intelligence Applications in Computer Engineering: Real Word AI Systems with Applications in eHealth, HCI, Information Retrieval and Pervasive Technologies,* pp. 3-24. Retrieved from http://dl.acm.org/citation.cfm?id=1566770.1566773

Kotsiantis, S. B. (2011). Decision trees: A recent overview. *Artificial Intelligence Review,* , 1-23. doi:10.1007/s10462-011-9272-4

Kotsiantis, S. B., Zaharakis, I. D., & Pintelas, P. E. (2006). Machine learning: A review of classification and combining techniques. *Artif.Intell.Rev., 26*(3), 159-190. doi:10.1007/s10462-007-9052-3

Liu, B. (2006). *Web data mining: Exploring hyperlinks, contents, and usage data (data-centric systems and applications)*. Secaucus, NJ, USA: Springer-Verlag New York, Inc.

Luo, P., Lin, F., Xiong, Y., Zhao, Y., & Shi, Z. (2009). Towards combining web classification and web information extraction: A case study. Paper presented at the *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining,* Paris, France. pp. 1235-1244. doi:10.1145/1557019.1557152

Minguillo, D., & Thelwall, M. (2011). The entrepreneurial role of the university: A link analysis of york science park. Paper presented at the *Proceedings of the ISSI 2011 Conference - 13th International Conference of the International Society for Scientometrics & Informetrics,* Durban, South Africa. pp. 570-583.

Park, H. W., & Thelwall, M. (2008). Link analysis: Hyperlink patterns and social structure on politicians' web sites in south korea *Quality & Quantity, 42*(5), 687 - 697. doi:10.1007/s11135-007-9109-z

Payne, N., & Thelwall, M. (2004). A statistical analysis of UK academic web links. *International Journal of Scientometrics, Informetrics and Bibliometrics, 8*(1)

Qi, X., & Davison, B. D. (2009). Web page classification: Features and algorithms. *ACM Comput. Surv., 41*(2), 1-31. doi:10.1145/1459352.1459357

Quinlan, J. R. (1993). *C4.5: Programs for machine learning*. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc.

Romero-Frías, E., & Vaughan, L. (2010). European political trends viewed through patterns of web linking. *Journal of the American Society for Information Science and Technology, 61*(10), 2109-2121. doi:10.1002/asi.21375

Romero-Frías, E., & Vaughan, L. (2012). Exploring the relationships between media and political parties through web hyperlink analysis: The case of spain. *Journal of the American Society for Information Science and Technology, 63*(5), 967-976. doi:10.1002/asi.22625

Smith, A. (1999). A tale of two web spaces: Comparing sites using web impact factors *Journal of Documentation, 55*(5), 577-592.

Smith, A. (2003). Classifying links for substantive web impact factors. *Proceedings of the 9th International Conference on Scientometrics and Informetrics,* Dalian, China,.

Stuart, D., Thelwall, M., & Harries, G. (2007). UK academic web links and collaboration - an exploratory study *Journal of Information Science, 33*(2), 231 - 246. doi:10.1177/0165551506075326

Thelwall, M. (2001). A web crawler design for data mining *Journal of Information Science, 27*(5), 319 - 325. doi:10.1177/016555150102700503

Thelwall, M. (2002a). The top 100 linked-to pages on UK university web sites: High inlink counts are not usually associated with quality scholarly content *Journal of Information Science, 28*(6), 483 - 491. doi:10.1177/016555150202800604

Thelwall, M. (2002b). Evidence for the existence of geographic trends in university web site interlinking. *Journal of Documentation, 58*(5), 563-574.

Thelwall, M. (2002c). An initial exploration of the link relationship between UK university web sites. *ASLIB Proceedings, 52*(2), 118-126.

Thelwall, M. (2009). Introduction to webometrics: Quantitative web research for the social sciences. *Synthesis Lectures on Information Concepts, Retrieval, and Services, 1*(1), 1-116. doi:10.2200/s00176ed1v01y200903icr004

Thelwall, M. (2001). Extracting macroscopic information from web links. *Journal of the American Society for Information Science and Technology, 52*(13), 1157-1168. doi:10.1002/asi.1182

Thelwall, M. (2003). What is this link doing here? beginning a fine-grained process of identifying reasons for academic hyperlink creation. *Information Research, 8*(3)

Thelwall, M. (2006). Interpreting social science link analysis research: A theoretical framework. *J.Am.Soc.Inf.Sci.Technol., 57*(1), 60-68. doi:10.1002/asi.v57:1

Thelwall, M., & Wilkinson, D. (2003). Graph structure in three national academic webs: Power laws with anomalies. *Journal of the American Society for Information Science and Technology, 54*(8), 706-712. doi:10.1002/asi.10267

Thelwall, M., & Wilkinson, D. (2004). Finding similar academic web sites with links, bibliometric couplings and colinks. *Information Processing & Management, 40*(3), 515-526. doi:10.1016/s0306-4573(03)00042-6

Vaseleiadou, E., & van den Besselaar, P. (2006). Linking shallow, linking deep. how scientific intermediaries use the web for their network of collaborators. *Internationl Journal of Scientometrics, Informetrics and Bibliometrics, 10*(1)

Vaughan, L., Kipp, M., & Gao, Y. (2007). Are co-linked business web sites really related? A link classification study *Online Information Review, 31*(4; 31), 440-450.

Vaughan, L. (2005). Mining web hyperlink data for business information: The case of telecommunications equipment companies *IN PROCEEDINGS OF THE FIRST IEEE INTERNATIONAL CONFERENCE ON SIGNAL-IMAGE TECHNOLOGY AND INTERNET-BASED SYSTEMS,* , 190; 190-195; -195.

Vaughan, L., & Thelwall, M. (2005). A modeling approach to uncover hyperlink patterns: The case of canadian universities. *Information Processing & Management, 41*(2), 347-359. doi:10.1016/j.ipm.2003.10.001

Vaughan, L., & Wu, G. (2004). Links to commercial websites as a source of business information *Scientometrics, 60*(3), 487 - 496. doi:10.1023/B:SCIE.0000034389.14825.bc

Wilkinson, D., Harries, G., Thelwall, M., & Price, L. (2003). Motivations for academic web site interlinking: Evidence for the web as a novel source of information on informal scholarly communication. *Journal of Information Science, 29*(1), 49-56. doi:10.1177/016555150302900105

MOVING FROM PERIPHERY TO CORE IN SCIENTIFIC NETWORKS: EVIDENCE FROM EUROPEAN INTER-REGIONAL COLLABORATIONS, 1999-2007 (RIP)

Lorenzo Cassi, Emilie-Pauline Gallié, Agenor Lahatte, Valérie Merindol

lorenzo.cassi@univ-paris1.fr
CES-University of Paris 1 and OST-Paris

emilie-pauline.gallie@obs-ost.fr
OST-Paris

agenor.lahatte@obs-ost.fr
OST-Paris

valerie.merindoli@obs-ost.fr
ESG Management School (Paris) and OST-Paris

Abstract
This paper provides an original framework for investigating scientific collaboration networks at European regional level. Is European scientific area an integrated system? Are the European regions organized around a core? Which are the determinants explaining the transition of region from periphery towards a core position? To answer these questions, the evolution of eight different scientific discipline networks from 1999 to 2007 is taken into account. For each of these networks, we perform a core-periphery and, given the transition matrices, we identify those regions which are able to move over time from a peripheral position to a more central one. A final exercise investigates the determinants of this transition. We conclude that fostering specialization in some discipline is the main explanation of why a regions is able to reach a more central position; moreover, once specialized in a domain, this could allow the region to benefit of a virtuous circle: increasing the overall number of publication and making a greater effort in science-based activities allow the region to become a core member also in other scientific domains.

Conference Topic
Collaboration Studies and Network Analysis (Topic 6).

Introduction
'Regional disparities' have been a main concern for European Union and consequently cohesion one of the key aim of its policy. This has been even more central since the recent enlargement to Eastern countries. These disparities have to do *not* only with economic income but also with innovation activities. The objectives of 2020 EU agenda in developing a European Research Area, on one

hand, ask for an increase of R&D expenditure in order to reach the threshold of three per-cent of GNP at the national level and, on the other hand, aim at fostering international collaboration in order to develop a more integrated research system.
This paper provides an original framework for the interpretation of the scientific collaboration between European regions. Its aim is to verify the existence of a core-periphery structure and, if it is the case, to identify the central and peripheral regions in eight different scientific disciplines and, finally, to investigate the main determinants explaining the capacity of a region to move from the periphery of the structure to its core.
To achieve this aim, we perform two empirical exercises focused on a subset of 178 European regions at NUTS 2 level in 21 countries. In the first exercise, we built up a network of scientific collaborations for each discipline in three different periods (i.e. 1999-2001, 2002-2004 and 2005-2007) and, for each of these networks, a core and a periphery are identified. In the second exercise, we examine the transition matrixes analyzing the determinants explaining why a region is able to reach a central position within the network.
The paper is organized as follows: the second section presents briefly the theoretical framework, the third one presents the data and discusses the main assumptions relate to network construction. The following ones present the two empirical exercises and the main results.

Theoretical background: collaboration network in Europe

The literature dealing with research collaboration among European regions is the main reference of our paper. This literature can be easily classified, among other criteria, according to data used as a proxy of collaboration *and* according to unity of analysis. Usually three kinds of data are analysed: research project (e.g. Framework Program), co-patent and scientific co-publication (for a review see Frenken et al., 2009). Concerning the unity of analysis, the literature focuses, on one hand, on individual actors (i.e. regions) analysing the position of each of them in collaboration networks or their collaboration pattern and, on the other hand, on the overall structure of network. The former is the local scale perspective, the latter the global one. Less studied is the intermediate-scale of EU research collaboration network. In this paper, using publication data, we investigate the meso-scale network feature known as core-periphery structure, which consists of a partition of network's nodes in a highly connected core and a sparsely connected periphery (Rombach et al, 2012).

Regional networks: data and method

In order to investigate the European scientific system we focus on the collaboration existing among regions (NUTS2) in eight broad scientific disciplines that are defined by OST (2010) as an aggregation of Thomson Scientific Categories. Data of co-publications among 248 NUTS2 regions (UE27) used and citations associated come from Web of Science (WoS) database, which contains the articles published in most journals covering all scientific fields. We

retrieve all scientific articles published between 1999 and 2007 and related to research collaborations among UE regions. A publication is considered to be research collaboration between regions if it contains at least two different institutional addresses corresponding to NUTS2 regions. The publications co-authored by intra-regional institutions are excluded. Our study is limited to the bilateral and multilateral co-authorship inter-regional and is in fact performed on three sets smoothed data corresponding to 3-years periods 1999-2001, 2002-2004 and 2005-2007. The analysis sample is otherwise restricted to regions having at least 500 publications all fields on average by period. That reduces our sample to 178 regions, 21 countries. Using this information, it is possible to build a network where the nodes are the regions and the links between them are given by the amount of their co-authored scientific publications. The analysis of this kind of network could give information on how the European system is structured.
For each region, we have calculated the following variables based on publication: the amount of publications in each discipline, the total amount of publication (all disciplines), a normalised specialisation index.[117] Moreover, for each region we have collected some economic data from the Eurostat website. In particular we downloaded for the period 1999-2007 for each region: the population (thousands of people); the human resources in Science and Technology, broadly defined (thousands of people); the GNP per capita relative to EU27 average, normalised to 100. Combing this data we define a new variable, labelled *Science-Based*, as the number of scientific publications (all disciplines) relative to number of people working in Science and Technology sectors. According to us, this measure of productivity can be also interpreted as a measure of how much the human capital of a region is dedicated to science related activities rather than more applied ones.

Core identification

The meso-scale perspective in Social Network Analysis literature has been mainly developed around the analysis of community structure (roughly equivalent to cluster) rather than the analysis of core-periphery. Core-periphery perspective is surprisingly few developed: the main reference is still the contribution of Borgatti and Everett (1999) and this is true both in terms of definition and calculation (to our knowledge, the only algorithm available is the one provided initially by UCInet).
According to them a "core-periphery model consists of two classes of nodes, namely a cohesive sub-graph (the core) in which actors are connected to each other in some maximal sense and a class of actors that are more loosely connected to the cohesive sub-graph but lack any maximal cohesion with the core" (Borgatti and Everett, 1999, p.377).

[117] The specialisation index is calculated as the ratio between the share of the region in one discipline and the share of the publication of the region in all fields. Moreover the index is normalised in the following way: $(index^2 - 1)/ (index^2 + 1)$. The range of normalized index is between -1 and 1: the index gets a value equal to 0 if a region has a share of publication in a discipline equal to its share in all fields.

A similar concept, that it is possible to find in the literature, is the *rich-club phenomenon* (Zhou and Mondargon, 2003) which consists in partition of network's nodes in a group of actors highly connected and the rest of network loosely connected. The main difference between the two definitions concerns who is connected with whom. In the core-periphery structure the peripheral ones are not connected (or very few) with each other and partially connected to the core, while in the rich-club structure there are not a priori assumptions on that but just a structural difference between actors in terms of degree (i.e. number of partners). Thus, in order to detect a rich-club structure, it is sufficient to analyse the degree distribution, and it is not required to look at the adjacency matrix as in core-periphery case. If a two-tier structure is identified, i.e. two different power laws characterising the degree distribution, then it is possible to infer the occurrence of a rich-club phenomenon. This latter represents a necessary condition in order to have core-periphery structure as defined by Borgatti and Everett.
Following the methodology developed in Zelnio (2012), we define a core and a periphery according to the rich-club phenomenon and, in order to do that, we jointly analyse the distribution of degree (partner) among regions and the regional distribution of number of article. If a region belongs to the rich-club according to both distributions, then the region is a member of the core (for more details on the methodology, see Zelnio 2012).
The following table displays the main results on core-periphery structure.

Table 1. Core: number of regions and share, by dsciplines

Disciplines	*1999-2001*	*2002-2004*	*2005-2007*
Fundamental Biology	40 (22.5%)	40 (22.5%)	47 (26.4%)
Medicine	44 (24.7%)	55 (30.9%)	50 (28.1%)
Applied Biology/Ecology	44 (24.7%)	51 (28.6%)	56 (31.4%)
Chemistry	51 (28.6%)	44 (24.7%)	55 (30.9%)
Physics	52 (29.2%)	49 (27.5%)	53 (29.8%)
Science of the Universe	46 (25.8%)	47 (26.4%)	50 (28.1%)
Engineering Sciences	29 (16.3%)	45 (25.3%)	50 (28.1%)
Mathematics	44 (24.7%)	51 (28.6%)	65 (36.5%)

The size of the core varies across disciplines and overt time. Moreover not all disciplines show the same increasing pattern. Indeed only Applied Biology, Engineering Sciences and Mathematics show this kind of pattern over the three periods, other as Fundament Biology and Science of the Universe are stable between the two first periods and increase later on. The other three disciplines report changes between the three periods of different sign. However for any discipline, the core of the last period is greater than at the beginning of the period under analysis.

Moving to the core

As Table1 shows, the size of core changes over times. This could happen because some regions move from the periphery to the core or the other way round. The following table presents the transition matrix which sums up the possible cases and reports the number of regions corresponding to each of them, without distinguished by discipline.[118]

The most interesting case concerns the regions which are able to move from the periphery to the core in a stable way, i.e. regions that have moved from the periphery to core between first and second period and that have been able to stay in the core over the last period. This latter is a subset of 97 cases listed in previous table, because among them we have 46 observations that corresponding to a region moving toward the core only between the second and third period. We prefer to exclude them by the following investigation because it could correspond to an "instable case" (i.e. a region moving back to periphery after one period in the core). However in future research we intend to extend the analysis also to these cases.

Table 2. Transition matrix over three periods for each region/discipline observation

		After	
		Core	*Periphery*
Before	*Core*	312 (21.9%)	21 (1.5%)
	Periphery	97 (6.8%)	966 (67.8%)

Note: there are 28 (2%) not corresponding to any listed cases, because instable.

Thus, in the following, we investigate the determinants of stable transition, i.e. 51 observations. In order to do that, we compare those regions with regions that have been in the periphery over the three periods (966 observations).[119] Table 3 reports the result of logistic regression analysing the probability the transition occurs according to two different specifications, respectively without and with interaction terms.

Table 3. Estimation of the probability to move form periphery to core (Logistic regression)

Variable	*Model 1*	*Model 2*
Population (thousands)	0.000644***	0.000677***
	[0.000134]	[0.000141]
GNP per capita (index)	0.0141**	0.0165**
	[0.00520]	[0.00601]

[118] That implies we have 1424 observation, i.e. 178 regions multiplied by 8 disciplines.

[119] Some observations are excluded from the following exercise because some missing information in Eurostat data.

Science-Based	0.1843***	0.1876***
	[0.0544]	[0.0561]
Discipline specialisation (Spec)	3.5394***	2.1896**
	[0.5389]	[0.6700]
Core in 1 disc (Icore1)	0.8086*	0.1359
	[0.3579]	[0.4570]
Core in 2-5 disc (Icore2)	2.1308***	1.7953***
	[0.3875]	[0.3711]
Core in 6-7 disc (Icore3)	2.2476***	3.6710***
	[0.5351]	[0.9715]
Spec*Icore11		2.6338^{+}
		[1.4631]
Spec*Icore12		2.1204*
		[0.9990]
Spec*Icore13		8.4114**
		[3.0286]
Fundamental Biology	-1.0949*	-1.6412**
	[0.4848]	[0.6184]
Medicine	0.0417	0.0280
	[0.4016]	[0.4077]
Applied Biology/Ecology	0.1845	0.2734
	[0.3671]	[0.3835]
Chemistry	-0.7838	-0.7853
	[0.4690]	[0.4996]
Physics	-1.6386**	-1.7229*
	[0.6247]	[0.6987]
Science of the Universe	-1.0653*	-0.9795
	[0.4596]	[0.4869]
Engineering Sciences	-0.1495	-0.2409
	[0.3776]	[0.4066]
Mathematics	Ref.	Ref.
Founding members	0.0262	0.0626
	[0.5166]	[0.6161]
Adhesion between 1953 and 1973	0.8341	0.8238
	[0.5139]	[0.6131]
Adhesion: 1974-1981	0.7348	0.7564
	[0.6265]	[0.7017]
Adhesion: 1982-1986	-0.0917	-0.1089
	[0.5723]	[0.6852]
Adhesion: 1987-1995	-3.0147	-2.9876
	[233.5]	[302.0]
Adhesion after 96	Ref.	Ref.
Intercept	-6.5688***	-6.4546***
	[0.9554]	[1.0615]

Number of observations: 918; Standard Errors in brackets; *** p<0.001, ** p<0.01, * p<0.05, $^{+}$ p<0.1

The size of the region (captured by the population) matters, as well as the relative GNP per capita does. The variable *Science-Based* (i.e. the number of scientific publications divided by to number of people working in Science and Technology sectors) affects positively the probability of moving to the core: the investment in human capital in science activities does pay off. The degree of specialisation in the discipline under analysis matters, as we can expect. A region needs to make an effort in a specific discipline in order to become a member of the rich-club regions. Moreover, being already a member of a core in other discipline plays a positive role. This effect is increasing in the number of core-disciplines as the three dummies Icore show.[120] According to the first specification (but results are more or less confirmed in the second one), only some disciplines do matter. Compare to the reference discipline (Mathematics), only Fundament Biology, Science of the Universe and Physics seem to affect the probability making the transition be less likely. This is not surprising given what we have observed in Table1: between the first two periods the size of the core of these three disciplines is stable or decrease when, in the same period, the core for Mathematics increases. The dummies relative to the date of adhesion to EU are not significant: membership age does not matter, suggesting that, if there is an effect of EU membership, this does not change as time goes by.

The second specification (Model 2) takes into account the interaction terms between specialisation index and the three dummies for core membership. In this case, the coefficient of specialisation in a discipline should be interpreted as the effect to being specialised when a region is not belonging to any core in other discipline. That means that one should compare the coefficient of the interaction terms with this one, in order to investigate the effect of being already a core member in other disciplines given a level of specialisation. What we observe is that this effect is increasing in the number of disciplines a region is already in the core (with exception of Icore2). Higher is the number of disciplines where a region is already core, lesser is the relative effort in a specific discipline that a region should make in order to became a member of the core also in that discipline. This implies that regions can benefit of a virtuous circle, if they are able to become a core region in at least one discipline. This result is partially smoothed if we look at the Icore coefficients (that should interpret as the effect when region is no specialized, i.e. specialisation index equals to zero). The effect is confirmed to be increasing, but for Icore1 that is not significant. This means that there is some mass critic effect: if a region does not make any effort in specialisation, it is not enough to be core in one other discipline in order to increase the probability to move toward a more central position. In order to benefit for spill-overs from other discipline, a region should be a member of the in at least two of them.

[120] ICore1 means that the region is already in the core of one other discipline, ICore2 means that is in the core in at least 2-5 other disciplines, and finally Icore3 means that the region is already in the core of other 6 or 7 disciplines.

References

Borgatti, S.P. and Everett, M.G. (1999). Models of core/periphery structures. *Social Networks*, 21, 375-395.

Frenken, K., Hardeman, S. & Hoekman, J. (2009). Spatial scientometrics: Toward a cumulative research program. *Journal of Informetrics*, 3, 222-232.

OST (2010), ***Indicateurs de sciences et de technologies*****, Economica, Paris.**

Rombach, M.P., M.A. Porter, J.H. Fowler and P.J. Mucha, (2012). *Core-Periphery Structure in Networks*, http://arxiv.org/abs/1202.2684

Zelnio, R. (2012). Identifying the global core-periphery structure of science. *Scientometrics*, 91, 601-615.

Zhou, S. and Mondragon, R.J. (2003). The Rich-Club Phenomenon in the Internet Topology, *IEEE Communications Letters*.

NANO-ENHANCED DRUG DELIVERY (NEDD) RESEARCH PATTERN FOR TWO LEADING COUNTRIES: US AND CHINA

Ying Guo,[1] Xiao Zhou,[2] Alan L. Porter[3] and Doug Robinson[4]

[1]*guoying_bit@163.com*
School of Management and Economics, Beijing Institute of Technology, Beijing (China)

[2]*belinda1214@126.com*
School of Management and Economics, Beijing Institute of Technology, Beijing (China)

[3]*alan.porter@isye.gatech.edu*
School of Public Policy, Georgia Institute of Technology, Atlanta (USA), and Search Technology, Inc., Norcross, GA (USA)

[4]*douglas.robinson@teqnode.com*
teQnode Limited, Paris (France)

Abstract
Nano-Enabled Drug Delivery ("NEDD") systems are rapidly emerging as a key nano application area. NEDD offers promise in addressing pharmaceutical industry challenges concerning solubility, cost-reduction, disease & organ targeting, and patent lifecycle extension. This study compares NEDD research patterns for the US vs. China by profiling data compiled by a multi-component search strategy in Web of Science. We present a range of analyses to address research activity trends, concentration differences, and collaboration networks corresponding to three characteristics of "New and Emerging Science & Technologies," for which NEDD represents a consequential case. It can help researchers and research managers understand the current status and future prospects of an emerging scientific or medical field. Such profiling of database search results can offer global insights to help discern main research trajectories, key players, and promising new shoots.

Conference Topic
Scientometrics Indicators (Topic 1) and Visualisation and Science Mapping: Tools, Methods and Applications (Topic 8).

Introduction

Nano-enhanced Drug Delivery (NEDD) systems

Nano-enhanced Drug Delivery (NEDD) systems seek to improve the release, distribution, absorption, and elimination of drugs. Traditional methods for administering drugs have relied on absorption in the digestive tract or skin or on injection (with manifold issues). We investigate new delivery methods that use

nanoparticles (e.g., lipid-based, polymer-based, proteins, dendrimers, etc.) to target specific organs or cell-types (Allen, 2004). These may increase drug effectiveness, both via "technical" and "social" effects – e.g., by controlling release one can reduce dose frequency and improve patient compliance. NEDD offers potential for treating chronic diseases and genetic disorders, and it has also been considered as a suitable substitute for conventional protein therapy.

China is on the rise in the drug delivery technology sector and is becoming an increasingly nimble competitor in the space. Chinese drug delivery companies are seeking to expand their opportunities into Europe and the US. One company, Lepu Medical Technology, uses its nanomaterial technology to make drug-eluting coronary stents, among other interventional cardiology products, and the company booked $120 million in 2011 revenue. Meanwhile, just as for other frontier technologies, the US has been the dominant leader in the area of biotech. Thus, with this context, it is important to underscore to what extent these two countries are progressing in this frontier technology -- NEDD.

Forecast Innovation Pathways (FIP) and Bibliometric Analysis

This NEDD study is part of a project seeking to develop methods to Forecast Innovation Pathways (FIP) for "New and Emerging Science & Technologies" (NESTs) (Porter et al., to appear). NESTs have great potential for innovation, but at the same time they are associated with great uncertainties. The nurturing of appropriate research avenues is crucial so that NESTs are developed along the most promising pathways, both in technological terms as well as towards addressing societal and economic problems or needs.

The project aims to develop a methodological framework and associated tools for analyzing NESTs to help policy makers and R&D managers to make better-informed decisions regarding innovation pathways. It combines empirical and expert knowledge of an emerging technology. The empirical work mainly seeks to extract intelligence from database search results about R&D activities, technological maturation, key players, and promising prospects for applications. This reflects a combination of bibliometrics and text mining (i.e., "Tech Mining" – Porter and Cunningham, 2005; Cunningham et al., 2006). The case studies in the project (one is NEDD) will be followed by expert interviews, first unstructured, then with q-method, plus a workshop with stakeholders to explore innovation pathways.

The development of new profiling and mapping techniques to characterize key actors and their interactions is crucial. Our hypothesis supposes that by understanding the various bodies of knowledge involved in a NEST, the key organizations, how they are related, and the visions they have constructed, analysts can grasp the diverse potential innovation pathways. Our approach aims to help analysts to identify previously hidden possibilities for connections among

new ideas, artifacts and actors relating to NEDD (Rammert, 2002) – hence trying to preserve diversity in order to avoid technological lock-in towards undesired applications (Stirling, 2007).

The content for this paper can be divided into four parts. First comes a general introduction. Contextual Framework and Research Approach follow. The third section presents bibliometric analysis results. The last section sums up and points out promising "next" research opportunities to pursue.

Contextual Framework & Research Approach

Data & Search strategy

Our study of NEDD originates from dissertation research in the Netherlands (Robinson, 2010) and from a separate study in support of the North Carolina Biotechnology Center's ("NCBC") efforts to stimulate innovation by matching research producers with companies having complementary drug delivery interests (Porter, 2010). Commencing in 2008, we devised a modular, Boolean, term-based search algorithm for NEDD, guided by knowledgeable colleagues in the US and Europe. We advanced a conceptual framework to approach NEDD, informed by various reviews and "foresight" pieces. This led us toward categorization to frame our current NEDD search (Zhou, 2013a).

Framework & Research questions

NESTs have some obvious characteristics. First, plenty of scientists believe in the future of any given NEST and apply themselves to advance it; so such technologies often show accelerating R&D activity and rapid development. Second, NEST R&D is often multidisciplinary or interdisciplinary, as is the case for nano science and engineering (Porter and Youtie, 2009). Third, because of the first two characteristics, NEST often calls for cooperative development, which could be among different countries, institutions, or researchers. When we explore the R&D activity for a given NEST, we address these three characteristics as indicators. In this paper, we apply them to NEDD.

As noted, our analyses of NEDD research activity presented in this paper focus on China and the US. In general, we would like to know, to what extent these two countries have developed competency in this high technology area? Is this technology providing a "window of opportunity" for these two countries to compete in the near future? Considering the three characteristics of NESTs mentioned, the study investigates in detail to what extent China and the US are asserting themselves to possibly establish dominance over NEDD applications to come. Specific questions that drive this paper are as follows:

- Research activity trend analysis (Rapid development): Does rapidly increasing publication activity mean "real" development?
- Research concentration difference analysis (Cross-disciplinary): How

scattered and different are Chinese and American NEDD research concentrations? This could help researchers locate and balance their research emphases.

- Research cooperation network analysis (Collaboration patterns): What does the network among countries look like, especially in terms of the positions of China and the US? To follow up, can we analyze and visualize the networks within countries to reveal different R&D mechanisms at work (and possibly suggest policy initiatives)?

The study also intends to demonstrate the importance of integrating bibliometrics and text analyses to generate informative innovation indicators, helping to construct a more informed picture of a country's performance.

Results for China and US NEDD Research Performance

Research activity trend analysis

Because WOS indexing of some 12,000 journals' content is done with some time lag, the data for 2011 and 2012 are incomplete. For trend analyses, we want to estimate the full activity for those years. We used total annual publication counts in recent years in WOS, expressly for its Science Citation Index (SCI) to normalize the NEDD data. We multiplied these ratios of expected/observed values to adjust the observed NEDD counts for these two years for each country to compare the research activity trend for China and the US.

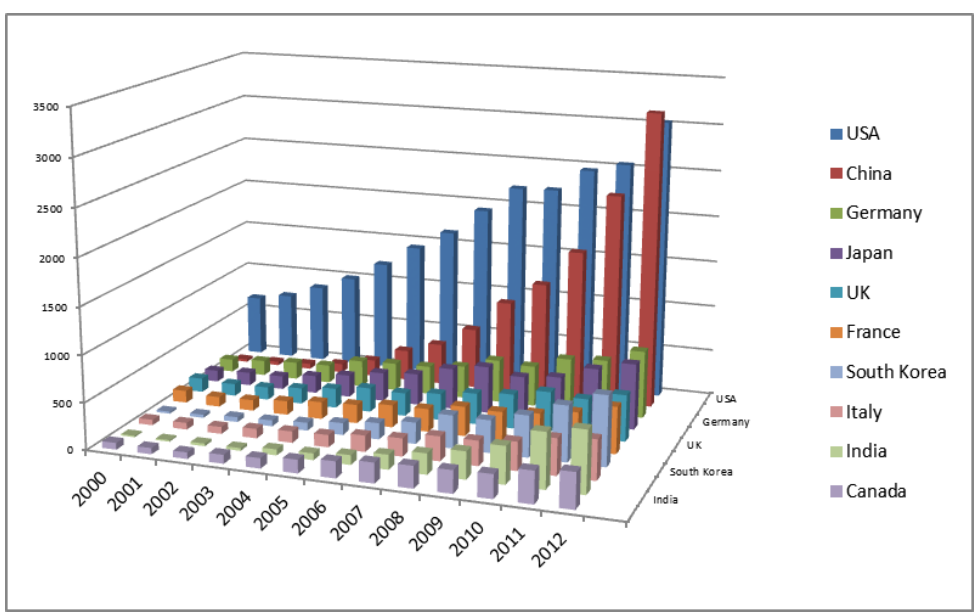


Figure 1. Activity Trends for the Top 10 NEDD Countries

It is surprising to observe China emerging as the most prolific in research publications in 2012 (although the data for 2012 are incomplete)! The picture has dramatically changed since 2001. In 2001, the US accounted for 45.5% of the NEDD papers, whereas Germany accounted for 10.4% of papers, and China for 2.5%. In 2011, China accounted for 23.5% of papers and US researchers authored or co-authored 26.2%, whereas Germany was considerably less visible, accounting for 5.8% of total papers (Figure 1).

For China, a steady rise in publication output has been observable particularly since about 2006. Taking 2001 as the base year, China's relative growth rate has been much higher than that of the US. Overall, for the aggregate publications from 2001 to 2012, China accounts for 10,110 NEDD papers (16.45% of the global total); the US, for 20,807 papers (33.85% of the total).

Citation measures provide a view of the reception of papers by the international community (Glanzel, 2008). However, any indicator based on citations received is strongly affected by the citation window. The larger volume of US papers leads to more citations than for China, and this is amplified by the fact that the US began publishing earlier within the 2000–2012 time period under analysis. The differences becomes less significant when this is normalized by number of papers and the number of years in which the paper is cited. We created Figure 2 to compare the citations/paper/year and the rate of uncited papers with the trend of total papers for both the US and China. Columns in the figures represent citation/paper/year for both countries and are oriented on the left Y axis; the line chart represents # uncited paper/total for each year for both countries and is oriented on the right Y axis.

Interestingly, we do not observe much difference in the corpus of papers that remain uncited, except for the year 2005. Considering the citations/paper/year, one important indication is how fast the papers of both countries are received by the international communityExcept for 2011 and 2012, citation for the US obviously progresses steadily, while for China, it looks erratic but possibly rising. Most tellingly, while China's citation rate lags behind that of the US, the gap seems to be moderate in recent years (note 2010 and 2011 especially). These bibliometric indicators imply that China's research is addressing important problems, advancing knowledge, and making researchers take note of that (Figure 2).

An examination of the top 100 most cited papers sheds further led on reception, revealing papers that attract the most attention. The top 100 highly cited papers wield major international impact. These papers may offer significant theoretical and/or experimental novelty to draw the attention of the research community. For this analysis, we fully credit a country for any paper on which it appears in one or more of the co-authors' addresses.

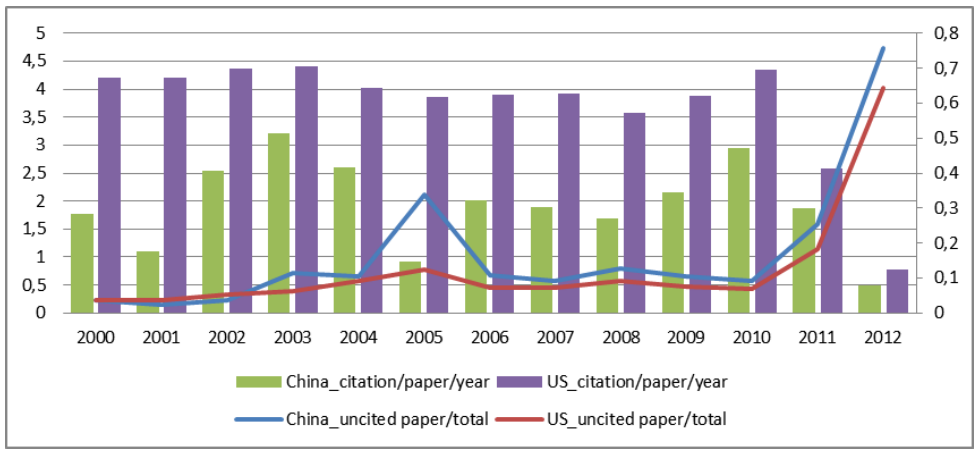


Figure 2. Citation trends for the US and China

The most cited papers present a rather different picture than average citation rates (Figure 2). We see only 2 Chinese papers among those top 100 highly cited papers, while 62 belong to the US. Also, interestingly, some countries appear with significant frequency in the top 100 highly cited papers – e.g. the Netherlands with 10 such papers (ranking 4th) – but remain absent in the top 10 highly active countries. This possibly means that the Netherlands NEDD community pays more attention to quality than quantity of research.

Research concentration difference analysis

We scan for hot topics within this area for both countries – i.e., topics within the domain that evidence increasing research attention in recent years. We identify hot topics by comparing the prevalence of key terms in a very recent period (here we use 2011–on) vs. an earlier period (here, 2000–2010). We examine 424 interesting, frequently occurring key terms in both periods. Table 1 lists the 20 topics that show the greatest increase in attention these past two years. Overall, the ratio of 2011–on to pre-2011 is 0.43. So, these terms are really "hot." For both countries, nano-related terms are hot, especially for the US. Also, we highlight 3 terms that appear in both lists, for China and the US. This illustrates the potential to identify respective research concentrations (to be probed further in consultation with several domain experts).

In our analyses, Principal Components Analysis (PCA) was applied to the most interesting (about 424) clumped term set (Zhang et al., under submission) to cluster these as potentially important factors for sub-systems and research concentration analyses for China and the US. We created a sub-dataset with US and China records, then made the factor map in Figure 3. We added the "country" field to each node to look for any big research concentration difference between the US and China. As for overall level, China accounts for 10,110 NEDD papers

(16.45% of the global total); the US, for 20,807 papers (33.85% of the total). The US publication number is almost double that of China. We have not shown the pull-down box for topical concentrations with similar ratios, but show those with interesting differences.

Table 1. Increasingly Popular NEDD Research Topics for US and China

US			
	A: #2011 on	B: #pre 2011	Ratio: A/B
siRNA delivery	166	102	1.63
Lipid nanoparticles	24	16	1.50
RAFT polymerization	24	18	1.33
photosensitizers	20	16	1.25
curcumin	28	23	1.22
silver nanoparticles	49	41	1.20
Nanoemulsion	27	29	0.93
alginate	19	21	0.90
Mesoporous silica nanoparticles	40	45	0.89
PLGA nanoparticles	46	54	0.85
Nanocarriers	93	110	0.85
radical polymerization	20	24	0.83
Gold nanoparticles	225	278	0.81
Hydroxyapatite	26	33	0.79
iron-oxide nanoparticles	135	175	0.77
Optical-properties	45	59	0.76
polymeric nanoparticles	83	110	0.75
mucoadhesion	6	8	0.75
Silica nanoparticles	53	71	0.75
Metal nanoparticles	29	39	0.74

China			
	A: #2011 on	B: #pre 2011	Ratio: A/B
Mesoporous silica nanoparticles	68	11	6.18
cell-penetrating peptides	32	8	4.00
siRNA delivery	108	29	3.72
cyclophosphamide	7	2	3.50
curcumin	22	9	2.44
resonance energy-transfer	24	10	2.40
Non-Hodgkins-lymphoma	7	3	2.33
signaling pathway	35	16	2.19
targeted delivery	64	30	2.13
Glioblastoma	21	10	2.10
In-vivo evaluation	27	13	2.08
Cell lung-cancer	37	18	2.06
Magnetic-resonance	22	11	2.00
photosensitizers	26	13	2.00
magnetic resonance imaging	51	26	1.96
Living cells	68	35	1.94
in-vitro evaluation	25	13	1.92
glioma	42	22	1.91
PROTEIN-KINASE	15	8	1.88
MRI	46	25	1.84

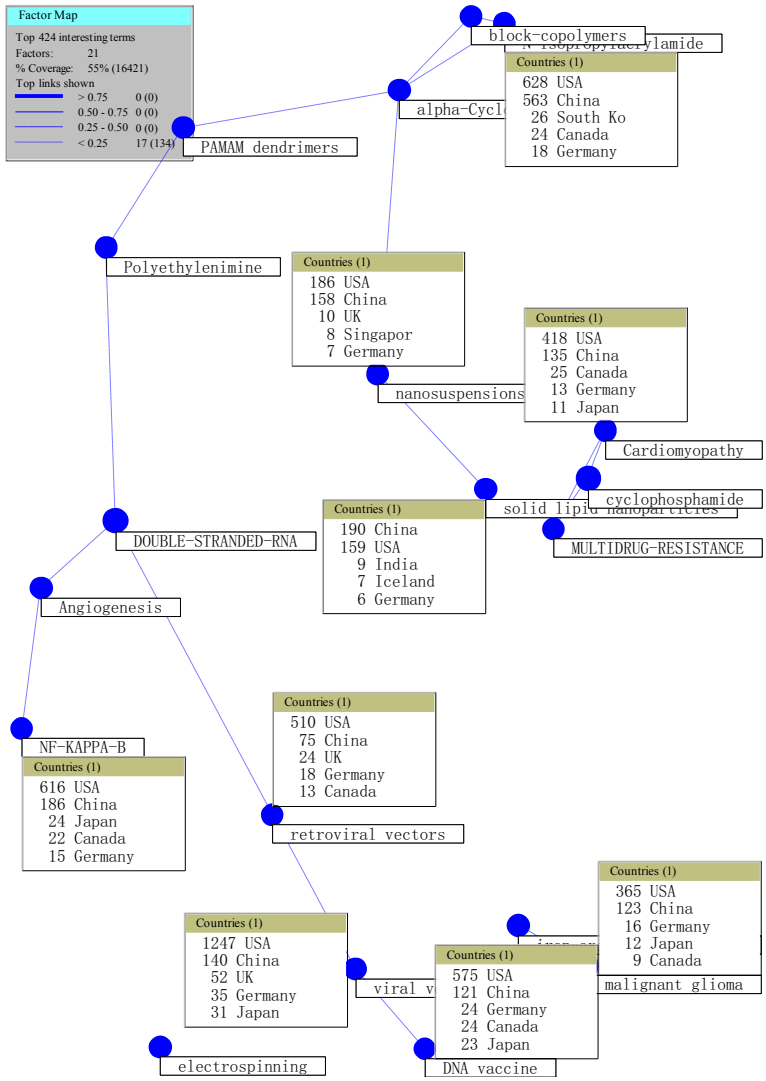


Figure 3. Factor map for US and China records

Looking at cardiomyopathy, retroviral vectors, viral vectors, NF-KAPPA-B, DNA vaccine and malignant glioma, China lags behind in terms of publications, especially for retroviral vectors and viral vectors. Such a big difference is surprising: viral or retroviral vectors are widely used in gene therapy, since they can directly deliver genetic material into cells. Nanosuspensions, N-isopropylacrylamide and solid lipid nanoparticles are related to easy delivery for certain drugs, using nano-size properties. China could compete with the US on number of publications. (We will probe further with the aid of experts in NEDD.)

Research collaboration network analysis

NEDD research is widely dispersed across different journals due to its interdisciplinary nature. More details of the activities of the two countries are highlighted in Table 2. This indicates some important aspects of these two countries' publication in journals with "Top 10" records in this field. Table 2 also reflects that international collaboration plays an important role in getting papers published in these journals, especially for *J. Virol* and *Mol. Ther*, for China. China shows a striking level of international collaboration for those journals. This analysis could also be conducted on the journals with a high impact factor, which could possibly show the publication quantity for each country.

Table 2. Journal Comparison for the US and China

Journal	# of records	China total	world share	international cooperation share	US total	world share	international cooperation share
Biomacromolecules	689	125	0.18	0.26	576	0.84	0.25
Biomaterials	1416	377	0.27	0.25	451	0.32	0.35
Gene Ther	670	30	0.04	0.33	442	0.66	0.30
Int. J. Nanomed	628	236	0.38	0.16	442	0.70	0.34
Int. J. Pharm	1447	284	0.20	0.18	357	0.25	0.30
J. Control. Release	1465	177	0.12	0.28	322	0.22	0.36
J. Mater. Chem	616	249	0.40	0.20	243	0.39	0.38
J. Virol	661	13	0.02	0.42	231	0.35	0.19
Langmuir	972	142	0.15	0.22	122	0.13	0.20
Mol. Ther	839	19	0.02	0.63	89	0.11	0.21

Further analysis for the top 100 highly-cited papers explores how closely countries cooperate (Figure 4). We also contrast research interests among countries based on these top papers (not shown here). Each node represents a paper in this map, and the links among them show how much they have in common. In Figure 4, the US is shown to collaborate with almost all the countries represented. We also learn from the research interests map analysis that, for the most part, the US shares interests with all the leading countries. All this uncovers the dominant status of US research in the global academy for NEDD. Further analysis could be made to dig into these papers to group and name them by using this map, with the aim of finding out the research areas that strongly attract attention.

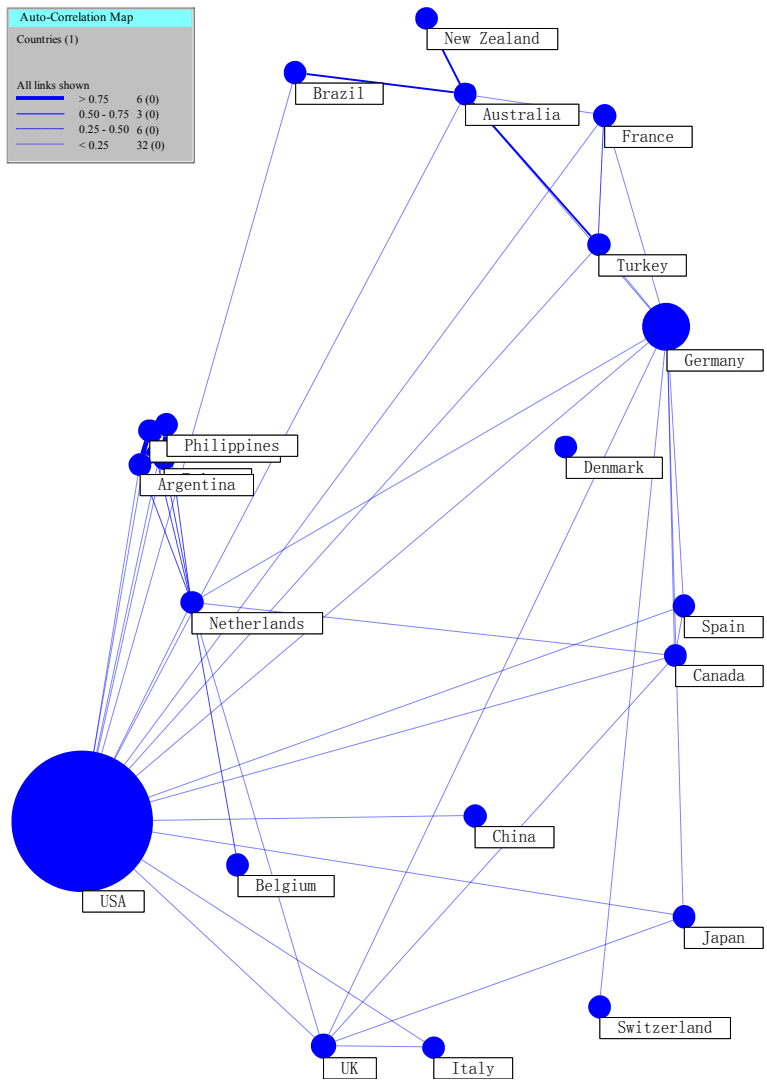


Figure 4. Cooperation Among Countries for the Top 100 Highly Cited Papers

Figure 5 highlights the links among the 20 top authors for both countries based on their co-authored publications. University and research institutes are dominating the network for both countries. Also cluster formation shows strong bearing on geographical location, e.g., the small group for Wuhan Univ., Sichuan Univ. of China. This formation may be due to sharing of capital-intensive instruments that are prerequisite for certain NEDD research. But the figure shows far fewer collaborative connections for the top 20 US authors, which is surprising. Some key differences should be noted. Chinese University-research institute linkages are very strong as almost all the institutes (majority being university) exhibit linkages with CAS, the Chinese Academy of Sciences (CAS represents over 100 research institutes, and over 400 S&T enterprises have been created by CAS).

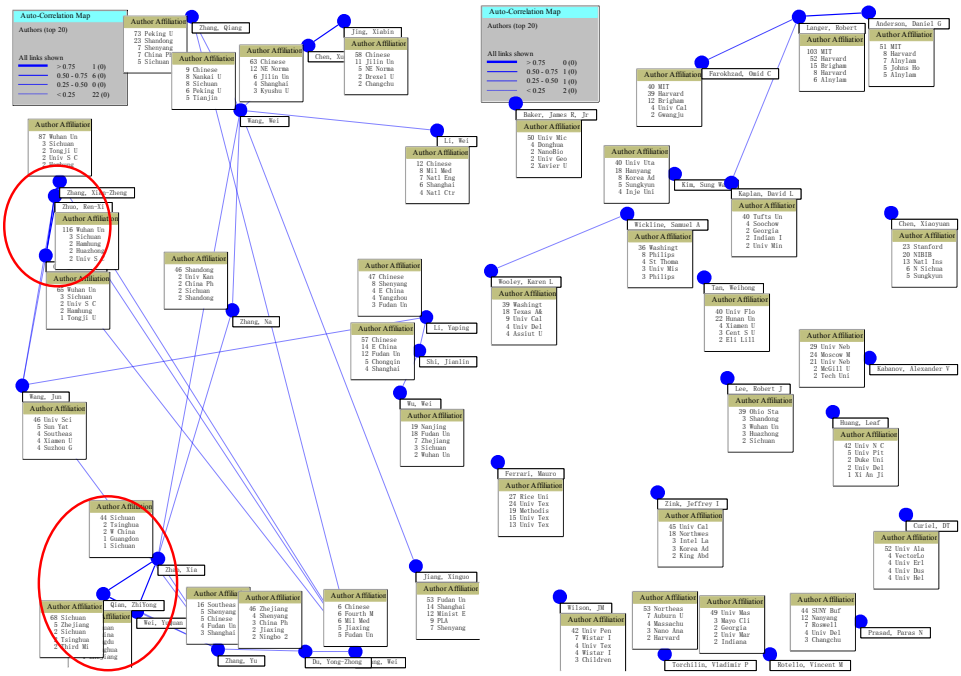


Figure 5. Top 20 authors for China and US

Discussion and conclusions

Nano-enhanced drug delivery (NEDD) systems seek to improve the release, distribution, absorption, and elimination of drugs. NEDD offers potential for treating chronic diseases and genetic disorders and has also been considered as a suitable substitute for conventional protein therapy. This paper conducts a comparative analysis of China vs. the US – two important players in the NEDD race – both to present results on this promising emerging technology and to consider ways to better perform such analyses.

After multiple iterations, we have developed a multi-module search strategy to construct an NEDD dataset from WOS, using the Georgia Tech (GT) "nano" (nanoscience, nanoengineering, nanotechnology, etc.) dataset with additional searches in the full WOS, led by our colleague, Xiao Zhou. Then we conduct several analyses to address research trends, collaboration pattern differences, and social network analyses concerning three characteristics of NESTs for which we aspire to forecast innovation pathways.

In the "research activity trend analysis," we compare both the activity and citation trend for the US and China. China's citation rate lags behind that of the US, but the gap has narrowed in recent years. Further, the trend has dramatically changed since 2001, as China has advanced notably in NEDD research. We then use "term

clumping" steps to clean and consolidate topical content in text sources. Analyzing the resulting key term set, we list the "hot" research topics for both countries. Using PCA to group key terms, we identify concentration difference for the two nations. The differences concerning viral or retroviral vectors are striking and merit reflection on R&D strategy. But China could compete with the US in terms of publication intensity for nanosuspensions, N-isopropylacrylamide and solid lipid nanoparticles, with implications for certain applications. Interestingly, the "research cooperation network analysis" shows that, although the US's international network spreads globally, internal collaboration seems somewhat limited.

The study strives to integrate bibliometrics and text analyses to generate informative innovation indicators, helping to construct a more informed picture of a country's performance. It further can help establish analytical steps to nominate and assess future innovation pathways ("FIP") for NEDD applications. That FIP process entails combining empirical findings with review and brainstorming by persons representing multiple stakeholder perspectives (Robinson et al., 2013).

For this paper, we focus on data from WOS. Since we know NEDD is heavily involved in medical science, we will also retrieve research publications from MEDLINE. We will then combine and compare with WOS, expecting about a 50% increase in R&D information to use in extended analyses. Next, as patenting is vital in pharmaceutical technology management, we will transfer and adapt the search logic to retrieve patent records in Derwent Innovation Index (DII).

Acknowledgements: This research draws on support from the National Science Foundation (NSF) Science of Science Policy Program – "Revealing Innovation Pathways" (Award No. 1064146) to Georgia Tech and has also been facilitated by NSF support through the Center for Nanotechnology in Society (Arizona State University; Award No. 0531194). The findings and observations contained in this paper are those of the authors and do not necessarily reflect the views of the National Science Foundation.

References

Arora, S.K., Porter, A.L., Youtie, J., and Shapira, P. (2013), Capturing new developments in an emerging technology: An updated search strategy for identifying nanotechnology research outputs, *Scientometrics*, DOI: 10.1007/s11192-012-0903-6.

Allen, T. M., Cullis, P.R. (2004) Drug Delivery Systems: Entering the Mainstream. *Science* 303, 1818-1822.

Cunningham, S.W., Porter, A.L.; Newman, N.C. (2006): Tech Mining Special Issue, Technology Forecasting and Social Change, Vol. 73 (8).

Glanzel, W. Seven myths in bibliometrics. About facts and fiction in quantitative science studies. In H. Kretschmer, & F. Havemann (Eds.), *Proceedings of WIS*

2008, Fourth International Conference on Webometrics, Informetrics and Scientometrics & Ninth COLLNET Meeting, Berlin: Humboldt University, 2008.

Porter, A.L., and Cunningham, S.W.(2005). *Tech Mining: Exploiting New Technologies for Competitive Advantage*. New York: Wiley.

Porter, A.L., and Youtie, J., "How Interdisciplinary is Nanotechnology? " *Journal of Nanoparticle Research*, vol. 11(5), 1023-1041, 2009.

Porter, A.L., Cunningham, S.W., and Sanz, A. (to appear). Extending the FIP (Forecasting Innovation Pathways) approach through an automotive case analysis, *Portland International Conference on Management and Engineering Technology (PICMET)*, San Jose, California, 2013.

Rammert, W. (2002) The cultural shaping of technologies and the politics of technodiversity. Sorensen, K.H. and Williams, R. (Eds.) *Shaping technology, guiding policy: concepts, spaces and tools*. Cheltenham, UK: Edward Elgar.

Robinson, D.K.R., Huang, L., Guo, Y., and Porter, A.L. (2013), Forecasting Innovation Pathways for New and Emerging Science & Technologies, *Technological Forecasting & Social Change*, 80 (2), 267-285.

Stirling, A. (2007). A general framework for analysing diversity in science, technology and society. *Journal of The Royal Society Interface*, 4(15), 707-719.

Zhang, Y., Zhou, X., Porter, A.L., and Gomila, J. (under submission), How to Combine Term Clumping and Technology Roadmapping for Newly Emerging Science & Technology Competitive Intelligence: The Semantic TRIZ Tool and Case Study, *14th International Society of Scientometrics and Informetrics (ISSI) Conference Proceedings*, Vienna, 2013.

Zhou, X., Porter, A.L., Robinson, D.K.R., and Guo, Y. (to appear, a), Analyzing Research Publication Patterns to Gauge Future Innovation Pathways for Nano-Enabled Drug Delivery, *Portland International Conference on Management and Engineering Technology (PICMET)*, San Jose, California, 2013.

Zhou, X., Porter, A.L., Robinson, D.K.R., and Guo, Y. (to appear, b), Patent and Publication Comparison for One Emerging Industries -- Nano-Enabled Drug Delivery, *14th International Society of Scientometrics and Informetrics (ISSI) Conference*, Vienna, 2013.

NANOTECHNOLOGY AS GENERAL PURPOSE TECHNOLOGY

Florian Kreuchauff[1] and Nina Teichert[2]

[1] *florian.kreuchauff@kit.edu* [2] *nina.teichert@kit.edu*
KIT Karlsruhe Institute of Technology, Dept of Economics, Chair in Economic Policy,
Kaiserstr. 12, D-76128 Karlsruhe, URL: wipo.econ.kit.edu

Abstract
Scientific literature postulates that nanotechnology is to be considered as general purpose technology (GPT), characterized by pervasiveness, high technological dynamism and the inducement of innovations within a variety of applications. We set out to not only further systematize existing approaches investigating nanotechnology's GPT traits based on patent applications, but to extend the analysis to academic publication data, in order to cover both knowledge creation and application development. By utilizing well established and consolidated indicators of GPT features, such as generality, diffusion, and forward citation rates, as well as contextualized technological coherence as a new weighted generality measure, we compare nanotechnology's research output to the ones of ICT as accepted GPT and of the combustion engine as a non-GPT, representing an upper and lower benchmark, respectively. Moreover, we add the EU27 as new institutional setting. Our results indicate that while nanotechnology is not as clearly perceptible a GPT as ICT is, the potential to develop as such and hence to become an 'engine of growth' is clearly given.

Conference Topic
Technology and Innovation Including Patent Analysis (Topic 5)

Introduction
Scholars emphasize that nanotechnology is not only *one* important but *the* general purpose technology (henceforth GPT) of the coming decade. Nanotechnology's versatile and interdisciplinary nature combines all classic basis technologies, promising revolutionary alterations of mankind's life, work, and perception of reality at all levels. GPT's sustainable economic surplus is created by the pervasive mutual inducements and complementarities of joint inventions in GPT and application sectors, yielding wide, continuously self-enhancing and accelerating impacts for the entire economy during whole eras (Bresnahan 2010). There is a vast literature examining whether past technologies are to be called a GPT, e.g. Lipsey et al. (1998) review potential candidates, Moser & Nicholas (2004) examine whether electricity was a GPT, & Jovanovic & Rousseau (2005) compare the impact of IT and electricity, to name just a few. However, it is considerably more difficult to investigate whether currently emerging technologies have the potential to become a GPT. The challenge arises because ex-ante even an exact definition of emerging technologies is difficult, without

even talking about ways to measure their impact. Nevertheless, conquering this bumpy road is important, because GPT's inherent innovation processes - though promising huge effects for economic growth - are subject to market failures and hence innovations are assumed to arrive too late and to a too little extent in terms of social welfare (Bresnahan & Trajtenberg 1995). Hence, if nanotechnology can be identified as young, but emerging GPT, sustainable policy implications can be derived in order to resolve, at least partly, the occurring market failures that hamper positive effects on productivity, enduring growth and prosperity.

We thus aim to contribute to the question, if nanotechnology is to be called an emerging GPT by validating that it features the three characteristics argued for as typical for general purpose technologies: *Pervasiveness* of use (1) is ensured by the generality of purpose, stemming from the possibility to arrange nanoscaled structures encompassing new material properties for literally countless applications in nanomedicine, atomically precise manufacturing, fuel cell electrocatalysis, organic photovoltaic cells and so on. The *scope for improvement* (2) in nanotechnology is provided by the possible reduction of size and costs, and increasing complexity. For instance, nanoapplications in semiconductor manufacturing technology have resulted in a remarkable reduction of processing size in recent years (Graham & Iacopetta 2009). Hints for nanotechnology to *spur innovation* (3) in application sectors are given by the existence of a nano-oriented value chain with basic, intermediate and downstream innovations (Youtie et al. 2008). Wang & Guan (2012) distinguish four stages within this value chain: nanomaterials, nanointermediates, nano-enabled products and nanotools. The relationship between electronic microscopy and nanotechnology sketches such possible value chains with inherent feedback loops exemplarily: R&D advances in instruments [e.g. scanning tunneling microscopes (STMs) / atomic force microscopes (ATMs)] actually opened the opportunity to conduct systematical research on the nanoscale, while advances in nanotechnology applied in such microscopes improved their capacities remarkably (Palmberg & Nikulainen 2006, Youtie et al. 2008). Thus, quality adjusted prices for ATMs and STMs declined, due to the application of nanotechnology. Moreover, in combination with the significant drop in scale enhancing the advances in semiconductors, this can also be instanced as evidence for innovational complementarities [combination of (2) and (3)]. We hence propose that nanotechnology is a general purpose technology and are subsequently testing the following hypotheses:

Hypothesis 1 *Nanotechnology is increasingly becoming a widely-used, pervasive technology.*

Hypothesis 2 *Nanotechnology exhibits scope for ongoing technological improvement.*

Hypothesis 3 *Nanotechnology increasingly spurs innovation in applications sectors.*

Methodology and Data

Previous Contributions and Systematic Extensions

In recent academic literature, nanotechnology has been progressively analyzed in order to identify economic trends attributable to its emerging nature. Various authors have contributed to the assembly of a holistic picture on nanotechnology's development, including Heinze (2004), who focuses on its worldwide expansion, Hullmann (2007), who examines data on markets, funding, companies, and patents and publications (concluding that nanotechnology easily has the potential to reach the level of the ICT's economic impact), Wong et al. (2007), who investigate the evolution of application areas, Meyer (2007), who emphasizes the integrating and field-connecting characteristics of instrumentation within nanotechnology, and Palmberg et al. (2009), who give a first broad overview on the development of nanotechnology. These lines of research already foreshadow nanotechnology being an emerging GPT. However, they neither formalize data analyses nor provide acknowledged measures for GPT traits, and thus lack a systematic investigation on this issue.

First *systematized* approaches to directly uncover GPTs (using patent data) were made by Hall & Trajtenberg (2006). They suggest measures for GPT attributes, such as a generality index, number of citations, and patent class growth, for patents themselves and for the patents that cite these patents. Alongside, basic approaches to investigate whether particularly *nanotechnology* might be a GPT were made by Palmberg & Nikulainen (2006). However, they do not yet apply those indicators proposed by Hall & Trajtenberg (2006) to test their hypotheses. These were adopted first by Youtie et al. (2008), who tested indicators for generality and highlighted evidence for nano being as pervasive as GPTs like ICT. Moreover, they developed new indicators for innovation spawning. Graham & Iacopetta (2009) also test for these two features, and Schultz & Joutz (2010) further deepened the topic, discovering a few very general emerging nano related fields with the potential for wide economic impact, and nano-fields that experience a more focused development path. Most recently, Shea et al. (2011) analyzed a sample of USPTO patenting activity of the first 25 nano-years, looking for early evidence that nanotechnology is a general purpose technology, assessing all three characteristics. Hence, first approaches to investigate GPT features within nanotechnology *systematically* have been developed. However, all of them were limited to patent applications and all investigating USPTO data.

We set out to not only further consolidate these existing approaches, particularly with respect to the indicators measuring the three GPT features, but we extend the analysis to publication data, in order to conquer both knowledge creation and application development. Moreover, although nano-activity has been subject to investigation by the OECD in recent years (Palmberg et al. 2009), to our knowledge there have not been any examinations of broadly accepted measures of GPT-characteristics within the EU27 yet. And finally, there has not been an answer to the need for distance measures between technology classes (Hall &

Trajtenberg 2006): Though pervasiveness constitutes the most highlighted GPT trait, the commonly stressed indicator, namely the so called generality index, suffers from the lack of distinction between closely related and very dissimilar technological fields. We thus not only utilize well-established and consolidated indicators of GPT features such as generality, diffusion, and forward citation rates, but add contextualized technological coherence as a new weighted generality measure, which has been demanded by Hall & Trajtenberg (2006), and with which we aim to complete the set of instruments on hand. Within all our analyses, we compare nanotechnology's research output to the ones of ICT as accepted GPT and of the combustion engine (henceforth CE) as a non-GPT, representing an upper and lower benchmark respectively.

Development tracking of GPTs with Patents and Publications

Patents, despite all difficulties that arise in their use and interpretation [see Porter et al. (2008) for an overview as well as Hullmann & Meyer (2003) and Huang et al. (2010) for a more detailed discussion on bibliometric issues concerned with nanotechnology], are widely accepted as proxy for innovative activity (Griliches 1990). Especially citation structures facilitate tracing knowledge flows [see Fischer et al. (2009), Bresnahan (2010), Jaffe et al. (1993), OECD (2009), Thompson (2006)]. Hence for the following analysis, data of nano-patents with priority application year between 1980 and 2008 were extracted from the 'EPO Worldwide Patent Statistical Database' (PATSTAT), version September 2010, and divided in samples including worldwide data and solely today's EU27. To identify relevant nano-patents by their titles and abstracts, a validated (evolutionary) lexical search strategy was used, based upon an approach of merging keywords proposed by Mogoutov & Kahane (2007), Glänzel et al. (2003) and Porter et al. (2008). CE and ICT patents were identified using search terms previously used in the literature: For CE, the IPC (International Patent Classification) class 'F02' was sufficient (Graham & Iacopetta 2009), whereas for ICT the search term was based on class definitions the IPC itself proposes. All patent queries are available upon request.

In addition, the considered nano-related **publications** are indexed in the Thomson-ISI WoS database. Again we refer to the period between 1980 and 2008. As well as with patents, a Boolean search term was used in order to identify nano-related publications by searching for certain keywords (and excluding others) in the topic of every paper. The search term is likewise based on the aforementioned combination of different search queries, but, due to technical restrictions, way shorter than the patent search term. A respective lexical CE query was developed by ourselves. For our GPT-reference ICT, we extracted all publications that were allocated in the Thomson ISI Subject Areas (SA) 'Computer Science' and 'Telecommunications', since an arguable description via keywords seems to be impossible for this field (Schmoch 2011, personal communication). As with patents, all queries are available upon request.

Results and interpretation

Pervasiveness (H1)

For a technology to be(come) pervasive, it has to be widely applicable already at an early stage of its development thereby using different diffusion channels. Finding evidence for nanotechnology being a future GPT thus includes finding linkages to a broad variety of different industries and technologies. Examining diffusion rates as one possible indicator of pervasiveness, one might consider the share of nano-patents / publications to total patents / publications in the respective portfolios of the most innovative firms and institutes, as diffusion is assumed to be fastest in these. Therefore, we apply this first quantitative measure exemplarily to the TOP25 firms in the European R&D Investment Scoreboard 2010 for patents and to the TOP25 publishing institutions in Europe (following WoS) for publications. In Figure 1, we depict the shares of ICT-, CE-, and nano-patents of the Top25 firms over the past three decades. As the trend indicates, the fraction of ICT-patents in innovative companies shows only a slight increase over the past 20 years (where one should not overrate findings in the last few data points: Interpreting patent developments demands caution regarding the last years, since patent acceptance takes its time. Due to this lag the last year in our sample is 2008, even though the database ranges till September 2010). It thus seems that there is a quite constant output rate of new codified applications in information and communications technology, so the growth follows a linear pattern. This is not only true for these 25 chosen companies, but for our observations of all patents as well.

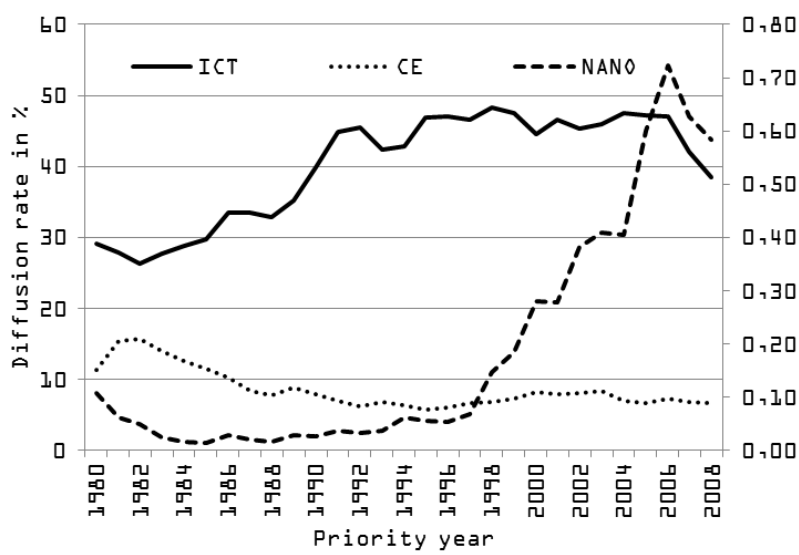


Figure 1. Patent Diffusion Rates of Top25 Firms in R&D, left axis: ICT and CE, right axis: nano

While the share of patents of our non-GPT proxy CE appears constant as well (around 7% percent for the last 20 years), the fraction of nano-patents seems to rise with a remarkable increase setting in about 1997. Nanotechnology inventions thus appear to gain in importance regarding their proportion of R&D-Output. But even in the observed companies with higher than average R&D intensity nanotechnology is still far away from outmatching the share of countable results in CE related research.

Scientific publications, though, are often associated with the more fundamental research, and nanotechnology evidences this quite clearly, as Figure 2 depicts. For the Top25 publishing institutions worldwide we observe shares of nano-related scientific literature around 6.5%, with an unbowed trend pointing to further growth in years to come. ICT shares of publications linger around 3%, with only a 1% increase in two decades. Hence ICT in general reveals a focus on applied research (as marked by patents), while nanotechnology is still primarily a matter of the scientific debate. Again, this is almost the same for the whole sample.

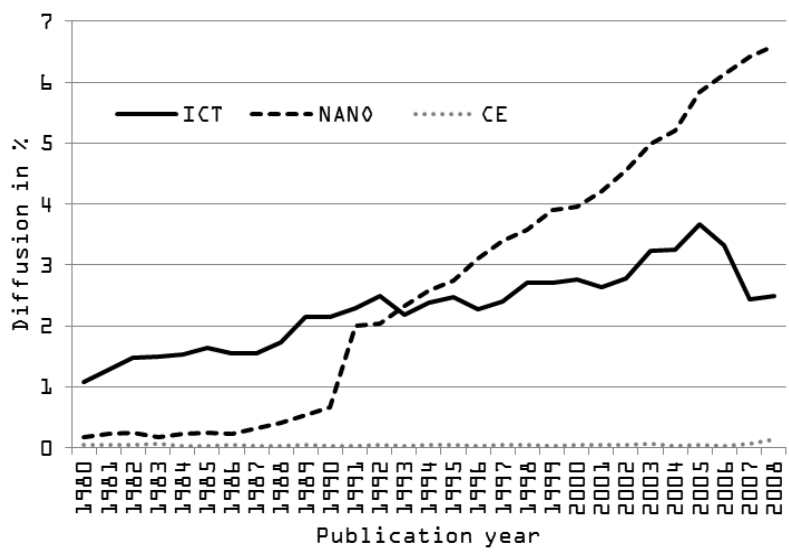


Figure 2. Diffusion Rates based upon publications of Top25 publishing institutions.

Already within their seminal paper, Bresnahan & Trajtenberg (1995) point to the possibility of identifying valuable inventions by patents that are cited by a wide range of different industries. To measure this, Trajtenberg et al. (1997) employed the Hirschman-Herfindahl index, which was further developed by Moser & Nicholas (2004) and Hall & Trajtenberg (2006) as generality index $G_i = 1 - \sum_j^{n_i} s_{ij}^2$, where s_{ij} denotes the percentage of citations received by patent i assigned to patent class j, out of n_i technological classes. If a patent benefited subsequent inventions in a wide range of technological fields, its generality index

will be close to one, whereas if most of its forward citations are concentrated in a small number of fields G_i will be close to zero. Correcting for the citation lag bias (small forward time windows associated with young and emerging technologies pose difficulties in calculating sensible generality indices, since not all the citations are yet observed, thus s_{ij} is biased downwards) is possible by using $\widetilde{G}_i = \frac{N_i}{N_i - 1} G_i$, where N_i denotes the total number of observed citations (Hall 2002). With respect to patents the generality index can not only be applied to IPC classes, but also be computed across technological fields in concordance with the International Standard Industrial Classification (ISIC) system. Such an aggregation generates less and broader defined classes, sharpening their distinctness, and yielding more meaningful generality indices. Thus, in our analysis, the underlying classes n_i do not represent 4-digit patent IPC classes, but 30 technological fields, in which these IPC classes are categorized in [following the NACE/ISIC Concordance developed by Hinze et al. (1997) according to OST/INPI/ISI - Observatoire des Sciences et Techniques / Institut Nationale de la Proprieté Industrielle / Fraunhofer Institut für System- und Innovationsforschung]. Calculations based upon IPC classes and their aggregation to 44 technological areas as developed by Schmoch et al. (2003) are available upon request. Figure 3 shows yearly average forward generality Indices of the Top10 cited patents according to the K30 technology classification (World data, EU27 available as well). Note that for CE we have calculated values for 5-year-intervals only, as we intended to keep the utilized amount of data at a reasonable level. Intermediate values are linearly interpolated. However, there is no reason to expect robustness problems by extending the data set.

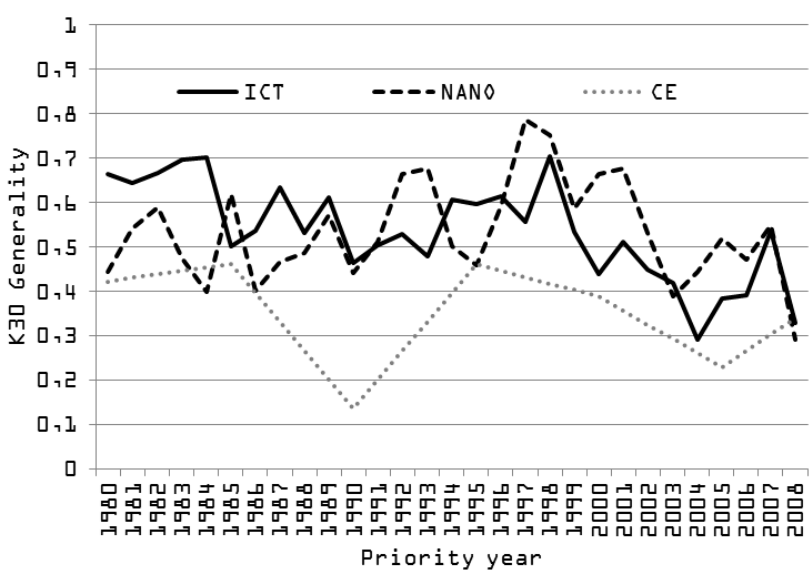


Figure 3. Forward Average Generalities of Top10 Cited Patents p.a. (K30).

Comparatively low generality indices seen in Figure 3 are explainable considering the fact that a smaller number of classes is taken into account: Less distinguishable classes entail smaller generality values, since all percentages of citations received by a patent are divided in fewer categories before their squares are summed up. The higher this sum becomes, the lower is the index. Fewer classes thus provide a higher accuracy of discrimination between pervasive technologies and those, of which the citation structure refers to a more limited number of fields. This is clearly to be seen in the figure: The average generality values of our lower benchmark CE are almost everywhere considerably smaller than those of ICT and nanotechnology. This holds true for the European sample.
The generality index is not restricted to patents. Publication data and the corresponding classification system of Subject Areas (SA) in Thomson ISI WoS can be used similarly. However, we do not show the results of our publication generalities here, since they offer little additional information: Classification within subject areas is subject to minor objectivity, which results in hardly distinguishable average generality indices.

The problem with generalities is best expressed by Hall & Trajtenberg (2006):
'[...] all of the generality measures suffer from the fact that they treat technologies that are closely related but not in the same class in the same way that they treat very distant technologies. This inevitably means that generality may be overestimated in some cases and underestimated in others. One suggestion for future research would be to construct a weighted generality measure, where the weights are inversely related to the overall probability that one class cites another class.'

We use a measure of technological coherence (TC) to approach this goal, which in our context will be defined as the extent to which inventions, i.e. patents, in a technological area share the same underlying knowledge. TC reflects the average relatedness of those classes, a patent is associated with, either because of being sorted in those classes or cited by them. Hence, to calculate the coherence of a patent portfolio, the degree of relatedness has to be determined for each pair of technology classes. Commonly, as e.g. in Breschi et al. (2003) and Leten et al. (2007), this is done using co-occurences of technological classes that are jointly associated to a patent. We will not recalculate the required relatedness matrix (with elements R_{ij}), but use the one constructed by Leten et al. (2007), which uses the OST / INPI / ISI concordance with 30 distinct tech fields. Following their citational approach, two technology classes are considered as technologically related if patents associated to one technology class often (i.e. more often than could be expected assuming random citation patterns) cite patents classified in the other technology class and vice versa. The patent-count weighted average relatedness $COH_i = \frac{\sum_{i \neq j} R_{ij} \times P_j}{\sum_{i \neq j} P_j}$, of technology i to all other technologies relevant in the considered year then leads to an overall coherence measure of (for example

nanotechnology) patents as a weighted average of all the COH_i measures: $COH = \frac{\sum_i P_i \times COH_i}{\sum_i P_i}$. We thus calculate the TC of (i) nano-patents applied for, and (ii) nano-patents citing patents, both within one year. TC can reasonably assumed to be higher, the more specialized a technological field is. New inventions in specialized fields are expected to be somewhat more *coherent* than are inventions in the field of a general purpose technology. By definition, GPT related inventions can be found in a wide range of application fields, and thus their TC is expected to be considerably smaller. We will employ this measure for the first time in this connection.

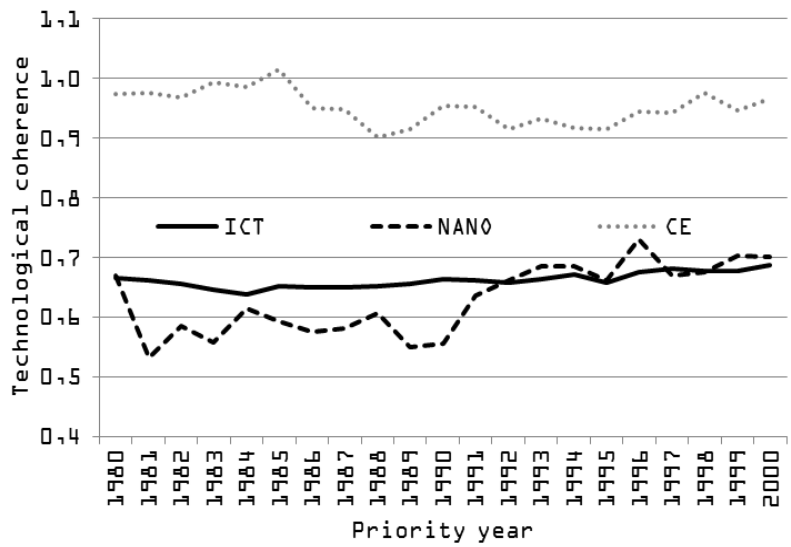


Figure 4. Technological Coherence of ICT-, Nano- and CE-Patents (World data).

Figure 4 shows the results for our TC-measure (i) based on world data. The GPT proxy ICT and nanotechnology shape a narrow side-by-side course with visible distance to the CE coherence values. To verify the significance of this offset we perform a two sample location t-test (available upon request). The results are robust when taking the technology classes of citing patents (ii) instead of the cited patents technology classes themselves, as well as when restricting the data to European patents (both available upon request). The measure is restricted to patents, since it relies on the relatedness matrix by Leten et al. (2007). Nonetheless, a similar matrix for publications might be constructed in further research.

With this new measure it becomes clear, that pervasiveness is undoubtedly much stronger for our ICT and the GPT candidate nanotechnology. Both show a visible distance to the lower benchmark technology CE, ICT with a smoother line due to

the clearer basis in the categorization system, nanotechnology with soft swings and a slight increase in coherence after 1990, the starting point of a significant rise in the number of nanotechnology patents, possibly due to a related small gain in concentration among technology classes.

Scope for Improvement (H2)

GPTs are improved continuously at every level of the value creation chain. Regarding nanotechnology and its potential to further reduce cost, size and enhance or even redefine material characteristics regarding stability, flexibility, abrasiveness, electrical properties and so on, two simple indicators shall illustrate the hitherto manifested scope for improvement.

With the first one we follow the suggestion of Palmberg & Nikulainen (2006) by observing the pure number of patents. We do not depict the results here for the sake of brevity, but as expectable, the number of nanotechnology patents has evolved noticeably over the past decades, though it is still far from reaching that of CE (not to mention ICT), a result strongly related to the contemporaneous lack of countable applications for the emerging technical feasibilities. As well as for diffusion rates, publications on the other hand again underscore the fundamental theoretical work that has been done for nanotechnology in the past 20 years. With the pure number of publications surpassing those of ICT at around the year 2000, nanotechnology has become *the* object of scientific interest of the new century. Nanotechnology's scope for ongoing improvements is thus unbowed, and there is little reason to expect any attenuation within the next years.

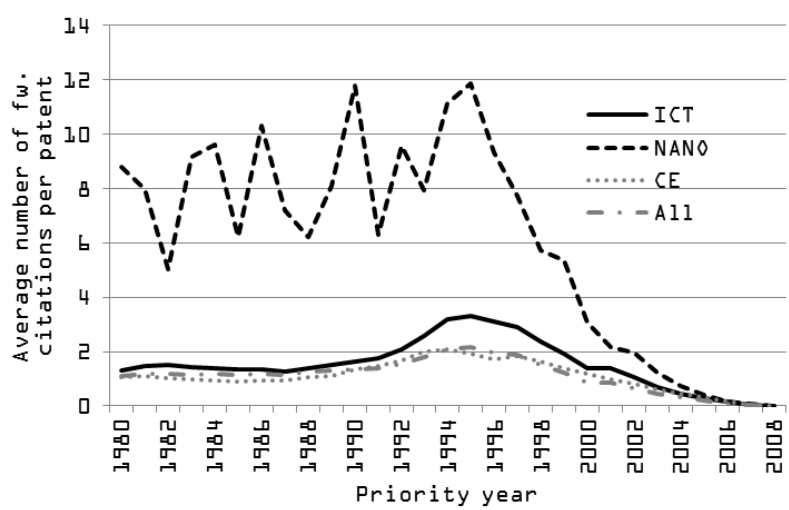


Figure 5. Forward Citation Rates ICT-, Nano- and CE-Patents (World data).

Our second indicator is based upon Schultz & Joutz (2010), who propose a later patent citing the original invention as an indicator for continual technological improvements. Following Hypothesis 2, nano-patents are hence expected to have many citations indicating a pattern of cumulative innovation (Hall and Trajtenberg 2006), an expectation which can easily be transferred to publications. In fact, we find nanotechnology producing patent citation rates even above those of ICT (and all patents worldwide, see Figure 5). A small absolute number of nano core patents produces comparably large numbers of references. These core technology founding patents seem to stem from outside Europe, since those nanotechnology patents we find in the European union have considerably smaller citation rates.
Publications are not affected that much by borderlines, and thus European *publications* again show high nano-related citation rates (due to database restrictions using scientific publications from WoS is considerably more difficult, which is why we limit ourselves to the European Union regarding publications). Again, visualized results are available upon request.

Innovation spawning (H3)

In the field of nanotechnology, innovation spawning can be found in the existence of nanoenhanced value creation chains, consisting of initial, intermediate, and downstream innovations. nanoscale structures (carbon nanotubes, quantum dots, fullerenes and so on) embodied in products with nanoscale features (coatings, optical components, or memory chips) and finally employed in a variety of final products (such as airplanes, computers, clothing, or pharmaceuticals) can be identified as such (Lux Research 2006, Youtie et al. 2008). In combination with technological dynamism, this characteristic is the main driver of innovational complementarities.
An increasing share of nano-inventions in overall patenting activity can be used as an indicator for the innovation spawning characteristic of nanotechnology. As for our Top25 firm sample, for the most part we find similar trends for the fraction of nano-, ICT-, and CE- patents worldwide, which is why we do not visualize them here. On account of this and for the sake of brevity again, we will focus instead on another indicator, namely the growth in nano-citing technological classes. If hypothesis 3 can be supported, nano-patents-citing tech classes are subject to a burst of innovations because they grow with the number of complementary goods developed (Hall & Trajtenberg 2006). A proxy for innovation spawning can hence also be the growth of technology classes (or subject areas with respect to publications) that harbour (nano-) citing patents / publications, as proposed by Hall & Trajtenberg (2006). Therefore we chose ten top citing patent classes (available upon request) according to their number of references, and ten subject areas according to a score system that accounts for the Top25 cited publications and the occurrence of their citations in these different subject areas. In the resulting diagram (figure 6) we cut the time before 1988, since we observed just a few classes in the beginning of nanotechnology's evolution, of which excessive

average growth would lead to the false impression that nanotechnology's trend was decreasing.

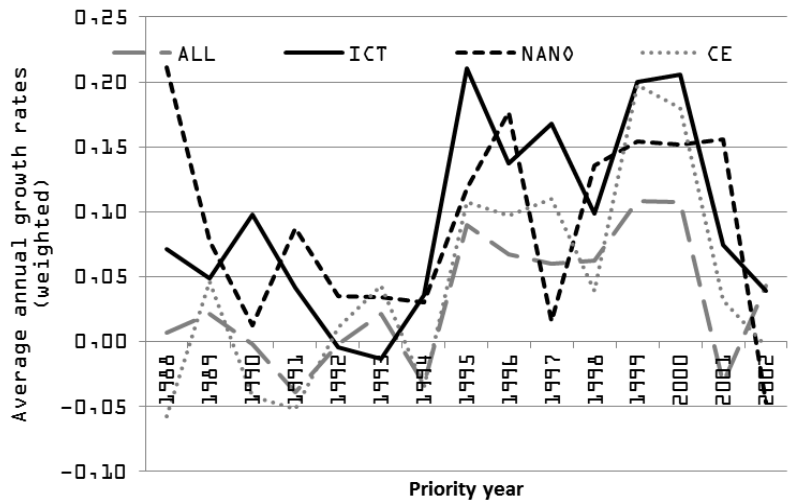


Figure 6. Growth of Top Citing Classes, ICT-, Nano- and CE-Patents (World data).

We cut above 2002 as well, since with declining overall citation rates (remember Figure 5) the average class growth becomes much less conclusive. Especially in highly complex technological areas (including undisputably our three compared technologies ICT, nanotechnology and CE) citations and therefore continual advancements take their time.

So while we are not willing to conceal an observed below-average class growth for all of those three technologies after 2002, one has to point out that the choice of classes is biased due to the declining observable citations. Thus with time, other classes might become more meaningful as predictor for an above-average class growth. Reselection of classes every year would lead to incomparability though, which is why being careful in interpreting the years after around 2000 is mostly without alternative.

For the remaining observation period nanotechnology and ICT both prove to be outstanding in their innovation spawning character. Almost without exception (1997 Nano, 1993 ICT) we find citing class growth to be above average. Admittedly, the lower benchmark CE does not perform too badly for this indicator as well, which is not surprising however: Though CE is not considered as GPT here, its ability to spawn innovation - even above average - within a less pervasive set of technological classes is unquestionable. Finally, regarding publications as supporting indicator, we do not observe significant above average growth rates. A straightforward explanation is yet to be found, but one might

guess that the method we chose to select the top ten subject areas (with the above-mentioned score system) could be responsible for that outcome.
Table 1 provides an overview of all our hypotheses, the analyzed measures and the corresponding results. Statements within the *Support* column reflect significant results from our t-tests regarding the visualized offsets (available upon request for all our measures) as well as our qualitative assessment with respect to level and trend. Keep in mind that the overall evaluation of the three GPT traits pervasiveness, scope for improvement, and innovation spawning ultimately relies on comparisons to the chosen benchmark technologies. Without these acknowledged counterparts and their scale function, any presented measure would lack relativization.

Table 1. Overview of Results Supporting the Hypotheses

Hypothesis	*Indicator*	*Result of nanotechnology*	*Support*
H1	Diffusion TOP25	PAT: way below ICT & CE, pos.trend	Weak
Pervasiveness		PUB: above ICT & CE	Strong
	Generality	Nano roughly between ICT & CE	Strong
	Techn. Coherence	Nano and ICT way below CE	Strong
H2	Increase in	PAT: way below ICT & CE, pos. trend	Medium
Scope for	Inventions	PUB: way above CE, surpassing ICT	Strong
Improvement	Forward Citation	PAT: way above ICT and CE/ALL (W)	Strong
		PUB: way above ICT and CE/ALL (EU)	Strong
H3	Diffusion	PAT: way below ICT, trends tw. CE (W)	Medium
Innovation		PUB: way above CE, surpassing ICT (EU)	Strong
Spawning	Citing Class	PAT: above average, similar to ICT	Strong
	Growth	PUB: average, below ICT, similar to CE	Weak

Conclusion

Stating that nanotechnology is widely considered as *the* general purpose technology of coming decades yields huge promises regarding consequent impacts on long term economic growth. GPT's three constituting characteristics pervasiveness, high technological dynamism and innovation spawning in various application fields have therefore been object of many studies. We contributed to this research by extending the underlying data to scientific publications, regarding Europe as examined region for the very first time, and adding up a new measure with technological coherence as demanded for. With an upper and lower benchmark technology, information and communication technology and the combustion engine, respectively, we provided comprehensive counterparts which proved to be useful comparisons.
The results indicate that nanotechnology evolves as GPT, as predicted by both scholars and practitioners. While it remains unclear if it yields similar potential as ICT has shown in the past two decades, nanotechnology's development regarding its unbowed continual advancement is undisputably as promising, as far as the data tell. Certainly, the incorporation of R&D expenditures representing the input

side would enable important insights when combining these two perspectives, offering explanations of macroeconomic growth already on the micro-level by investigating incentives and their interdependencies. This enrichment should facilitate the political discussion regarding emerging GPTs, especially as soon as country-level data reveals catch-up potentials. Moreover, by adding impact measures of national (or for instance European) and institutional technological leverage capabilities, inference statistics could provide a more holistic view on nanotechnology and even more, on GPTs altogether.

Acknowledgments

The authors wish to thank the three anonymous reviewers for their valuable comments and suggestions that have led to the improvement of this article. Parts of this work, including all figures and tables, can be found in: Teichert, N. (2012). Dissertation zur Erlangung des akademischen Grades eines Doktors der Wirtschaftswissenschaften. *Innovation in general purpose technologies: How knowledge gains when it is shared*. Karlsruhe: KIT Scientific Publishing.

References

Breschi, S. & Lissoni, F. (2001). Knowledge spillovers and local innovations systems: A critical survey. *Industrial and Corporate Change*, 10, 975-1005.

Breschi, S., Lissoni, F. & Malerba, F. (2003). Knowledge-relatedness in firm technological diversification. *Research Policy*, 32, 69-87.

Bresnahan, T. (2010). General purpose technologies. In B. Hall & N. Rosenberg (Eds.), *Handbook of Economics of Innovation*, Vol. 2 (pp. 761-791). Amsterdam: Elsevier.

Bresnahan, T. & Trajtenberg, M. (1995). General Purpose Technologies: 'Engines of growth'?. *Journal of Econometrics*, 65, 83-108.

David, P.A. (1991). Computer and dynamo: The modern productivity paradox in a not-too distant mirror. *Technology and Productivity: The Challenge for Economic Policy* (pp. 315-347). Paris, OECD.

Fischer, M., Scherngell, T. & Jansenberger, E. (2009). Geographic localisation of knowledge spillovers: Evidence from high-tech patent citations in europe. *Annals of Regional Science*, 43, 839-858.

Glänzel, W., Meyer, M., du Plessis, M., Thijs, B., Magerma, T., Schlemmer, K., Debackere, K. & Veugelers, R. (2003). Nanotechnology Analysis of an emerging domain of scientific and technological endeavour. *Steunpunt O&O Statistieken*.

Graham, S. & Iacopetta, M. (2009). *Nanotechnology and the emergence of a General Purpose Technology*. Retrieved January 21, 2013 from: http://papers.ssrn.com/sol3/papers.cfm?abstract_id=1334376

Griliches, Z. (1990). Patent statistics as economic indicators: A survey. *Journal of Economic Literature*, 28, 1661-1707.

Hall, B. (2002). A note on the bias in the herfindahl based on count data. In A. Jaffe and M. Trajtenberg (Eds.), *Patents, Citations, and Innovation* (pp. 454-459). Cambridge, MA: MIT Press.

Hall, B. & Trajtenberg, M. (2006). Uncovering GPTs with patent data. In B. Hall and M. Trajtenberg (Eds.), *New frontiers in the economics of innovation and new technology: Essays in honor of Paul A. David*. Northhampton, MA: Edward Elgar.

Heinze, T. (2004). Nanoscience and nanotechnology in Europe: Analysis of publications and patent applications including comparisons with the United States. *Nanotechnology Law & Business*, 1(4), 427-445.

Hinze, S., Reiss, T. & Schmoch, U. (1997). Statistical analysis on the distance between fields of Technology. Report for European Commission TSER Project.

Huang, C., Notten, A. & Rasters, N. (2010). Nanoscience and technology publications and patents: A review of science studies and search strategies. *Journal of Technology Transfer*, 36, 145-172.

Hullmann, A. (2007). Measuring and assessing the development of nanotechnology. *Scientometrics*, 70(3), 739–758.

Hullmann, A. & Meyer, M. (2003). Publications and patents in nanotechnology. An overview of previous studies and the state of the art. *Scientometrics*, 58(3), 507–527.

Jaffe, A., Trajtenberg, M. & Henderson, R. (1993). Geographic localization of knowledge spillovers as evidenced by patent citations. *The Quarterly Journal of Economics*, 108(3), 577–598.

Jovanovic, B. & Rousseau, P. (2005). General purpose technologies. In P. Aghion and S. Durlauf (Eds.), *Handbook of Economic Growth*, Vol. 1B (pp. 1181-1224). Amsterdam: Elsevier.

Leten, B., Belderbos, R. & van Looy, B. (2007). Technological diversification, coherence and performance of firms. *Journal of Product Innovation Management*, 24(6), 567-579.

Lipsey, R., Bekar, C. & Carlaw, K. (1998). What requires explanation?. In E. Helpman (Ed.), *General Purpose Technologies and economic growth* (pp. 15-54). Cambridge, MA: MIT Press.

Lux Research (2006). *The Nanotech Report 2006*, 4 edn. New York: Lux Research.

Meyer, M. (2007). What do we know about innovation in nanotechnology? Some propositions about an emerging field between hype and path dependency. *Scientometrics*, 70, 779-810.

Mogoutov, A. & Kahane, B. (2007). Data search strategy for science and technology emergence: A scalable and evolutionary query for nanotechnology tracking. *Research Policy*, 36, 893-903.

Moser, P. & Nicholas, T. (2004). Was electricity a general purpose technology? Evidence from historical patent citations. *American Economic Review, Papers and Proceedings*, 94(2), 388-394.

OECD (2009). Patents as statistical indicators of science and technology. *OECD Patent Statistics Manual*. Paris: OECD.

Palmberg, C., Dernis, H. & Miguet, C. (2009). Nanotechnology: An overview based on indicators and statistics. *OECD STI Working Paper 2009/7*.

Palmberg, C. & Nikulainen, T. (2006). Nanotechnology as a general purpose technology of the 21st century? - An overview with focus on Finland. *DIME Working Paper Series*.

Porter, A., Youtie, J., Shapira, P. & Schoeneck, D. (2008). Refining search terms for nanotechnology. *Journal of Nanoparticle Research*, 10, 712-728.

Schmoch, U., Laville, F., Patel, P. & Frietsch, R. (2003). Linking technology areas to industrial Sectors. Final Report to the European Commission, DG Research.

Schultz, L. & Joutz, F. (2010). Methods for identifying ermerging general purpose technologies: A case study. *Scientometrics*, 85, 155-170.

Shea, C., Grinde, R. & Elmslie, B. (2011). Nanotechnology as general purpose technology: Empirical evidence and impliations. *Technology Analysis & Strategic Management*, 23(2), 175-192.

Thompson, P. (2006). Patent citations and the geography of knowledge spillovers: Evidence from inventor- and examiner-added citations. *The Review of Economics and Statistics*, 88(2), 383-388.

Trajtenberg, M., Jaffe, A. & Henderson, R. (1997). University vs. corporate patents: A window on the business of innovations. *Economics of Innovations and New Technology*, 5(2), 19-50.

Wang, G., Guan, J. (2012). Value chain of nanotechnology: A comparative study of some major players. *Journal of Nanoparticle Research*, 14(2), 1-14.

Wong, P., Ho, Y. & Chan, C. (2007). Internationalization and evolution of application areas of an emerging technology: The case of nanotechnology. *Scientometrics*, 70(3), 715-737.

Youtie, J., Iacopetta, M. & Graham, S. (2008). Assessing the nature of nanotechnology: Can we uncover an emerging general purpose technology?. *Journal of Technology Transfer*, 33, 315-329.

NEVIEWER: A NEW SOFTWARE FOR ANALYZING THE EVOLUTION OF RESEARCH TOPICS

Qikai Cheng[1], Xiaoguang Wang[1], Wei Lu[1], Shuguang Han[2]

[1] {chengqikai0806@gmail.com, whu_wxg@126.com, reedwhu@gmail.com}
School of Information Management, Wuhan University, Wuhan, China, 430072

[2] shh69@pitt.edu
135 N Bellefield Avenue, Pittsburgh, PA, United States, 15213

Abstract
This paper proposes a new research frame for analysing the evolution of research topics in a discipline based on co-word network analysis. A new software was introduced, i.e. the NEViewer, in which we present three key features that are distinct from other science mapping tools: (a) a powerful analysing module within a longitudinal framework; (b) the use of several network community evolutions analysing algorithms; (c) revealing the macroscopic shifts and microcosmic details of evolution based on alluvial diagram and colored network. Our experimental analysis using five computer science conferences dataset show that the NEViewer is effective and reliable; and the research process using co-word network analysis in disciplines is also feasible.

Conference Topic
Collaboration Studies and Network Analysis (Topic 6) and Visualisation and Science Mapping: Tools, Methods and Applications (Topic 8).

1. Introduction
Methodologies and techniques developed in complex networks and information visualization fields provided opportunities for researchers in information sciences to detect and visualize the latent knowledge structure of research topics. Recently, the knowledge map and knowledge networks (Börner, Chen, & Boyack, 2005; Boyack, Klavans, & Börner, 2005; Chen, 2005; Leydesdorff & Rafols, 2008) have gained much attention in academia, in which researchers combined techniques from both two fields and aimed to uncover the latent scientific topic structures. After mapping topics and topic bursts in PNAS (Mane & Börner, 2004), the authors concludes that the knowledge network structure is particularly important in discovering hidden human knowledge, detecting hot topics and identifying research trends. As one of the most important knowledge networks, co-word network has been identified as one of the most important knowledge map and the alternatives of the citation networks or co-citation networks. Comparing to traditional method, it is even better in research trend detection because the

formation of citation networks and co-citation networks usually required too much time (Börner et al., 2005; Chen, 2005; Mane & Börner, 2004).
Co-word network is one type of networks; therefore, methods and algorithms for network analysis can be applied to co-word network. In many real natural/social networks, the roles of nodes in the networks are not homogenous. One common characteristic of social network is that it has latent community structure (Ding, 2011): a group of nodes may have more close relationships with each other than with the rest of nodes. The same phenomena were also detected in the co-word networks. The latent structure in the co-word networks has certain connections with the existing discipline structure. Keywords in different communities can be mapped into different disciplines, research topics. The evolution of co-word networks actually reveals the topic evolution process (Wang, Jiang, & Li, 2010).

Software packages such as CiteSpace and SciMat have been developed in detecting and visualizing topic evolution and knowledge mapping. However, CiteSpace only focused on the citation networks, which cannot provide real time topic evolution analysis because there is a delay in forming the citation networks. SciMat cannot provide sufficient details of network evolution, and their visualization was not easy to use. Therefore, we developed the NEViewer, a software toolkit that is able to detect and visualize the network evolution and topic evolution, and provide both micro-level and macro-level network analysis. In the following section 2, we will introduce the software architecture and related algorithms which are implemented in our software. Section 3 started with the introduction of the NEViewer, and then we provided a case study in NEViewer using five conferences dataset. The experimental results actually demonstrated the effectiveness of using our software. This paper concluded with the summarization of the advantages and disadvantages of the NEViewer.

2. Methodology

Scientific publications are important media for researchers to publish their research outcomes (Cobo, López-Herrera, Herrera-Viedma, & Herrera, 2010; Wang et al., 2010). Usually, the authors are required to provide several carefully selected keywords to represent the main research topic of the paper. Therefore, keywords are considered as important and controlled vocabularies for research and study (Lin, Zhang, Zhao & Buzydlowski, 2012). The co-keyword networks are also valuable and are able to reveal the latent relationship among research topics and research domains. By simply adding the temporal information, we will be able to investigate the evolution of co-word networks, and further map the evolution to the topic level.
The keyword network based topic evolution analysis can be divided into three phases: the pre-processing step, the topic identification step and the sequential data construction step. First of all, we need to pre-process the raw data and convert them into a sequence of temporal co-word networks. We assign one time stamp for each of those networks based on the publication date. Then, the

community detection algorithms are adopted to uncover the latent community structures within the co-word networks in each time stamp. Since we are focusing on the high-level topic evolution, each community is assumed to be a topic and a representative keyword will be assigned to represent this topic. The last step is to map the communities (topics) from different time stamps and generate a final sequential evolution of those topics. The NEViewer then visualized the sequential evolution of those topics.

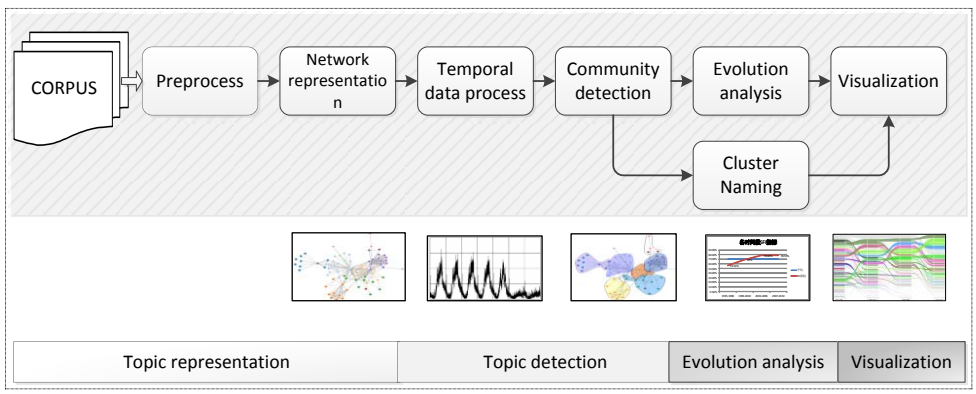


Figure 1. The framework of NEViewer's methods

2.1 Community Detection in Co-word Networks

There are two main approaches in identifying the latent structures within a given dataset: the network topology based approach and the content-based topical analysis approach (Ding, 2011). The first approach is based on existing well-investigated theories and methods in *Graph Theory*. A lot of algorithms were proposed in the last decade by computer scientists and physicists. Modularity Maximization is a widely adopted algorithm for community detection (Newman, 2004). Its basic idea is to traverse all possible community divisions and choose the division which can maximize the network modularity. Modularity is a metric for measuring whether the community results far away from the random assignments. The high modularity implies networks have dense connections within communities but sparse connection between communities. Although the initial algorithm was designed for un-weighted networks but it can also be extended to weighted networks, as suggested by its author (Newman, 2004). One problem for the Modularity Maximization method is that each node can only belong to one community, which may have conflictions with human's intuition. For example, one information retrieval (information retrieval community) expert can be the same time a social network (social network community) expert. Motivated by this idea, Palla, Derényi, Farkas and Vicsek (2005) proposed a K-cliques algorithm based approach, in which each node was able to be assigned to multiple communities. The authors further built the tool CFinder for visualization. Ball, Karrer and Newman (2011) proposed another approach based on a well-

known topic modelling algorithm: the probabilistic latent semantic analysis (PLSA). The output of their algorithm for each node's community detection result was probability distributions over all communities. By comparing with several existing approaches, Lancichinetti and Fortunato (2009) concluded that both the information-theory based approach and Blondel's approach (Blondel, Guillaume, Lambiotte, & Lefebvre, 2008) are superior in their evaluation. McCain (2008) adopted the second approach and find the effectiveness of content-based topical analysis in citation networks. Later, Wallace, Gingras and Duhon (2009) found that applying community detection algorithms in research domain analysis is reasonable, and even superior in uncovering more detailed information about the knowledge structures.

2.2 Representative Node Finding

Given that the latent community structure has been detected, we still need to find the representative topics for each community, i.e. the representative node finding problem. The nodes in co-word networks are the paper keywords (also the research topics), identifying the representative topics for each community can be converted into the searching of the most representative nodes. The keywords of those representative nodes can be used to represent the community topics.

Guimerà, Sales-Pardo and Amaral (2006) classified the roles of nodes into several categories: the provincial node, the connector hub node, the peripheral node, and etc. They further proposed to use the Z-value for determining the roles. The Z-value measures the local structural position of each node in the whole network. It is defined in the following formula, in which k_s^i represents the number of links between node *i* and community s; s_i indicates the community that node *i* belongs to; [illegible] denotes the average number of nodes in community s. A higher Z-value implies a higher closeness of node *i* and the other nodes in community s_i. Based on the suggestions in Guimerà, Sales-Pardo and Amaral (2006) and our experiences, each detected community can be represented by one or more nodes who are in the community and whose Z-values ≥ 2.5.

$$z_i = \frac{k_{s_i}^i - < k_{s_i}^j >_{j \in s_i}}{\sqrt{< (k_{s_i}^j)^2 >_{j \in s_i} - < k_{s_i}^j >^2_{j \in s_i}}}$$

2.3 Community Evolution Analysis

The co-word networks in different timestamps are usually unstable: the network density, the number of communities and the size of each community are all likely to change in different time stamps. The evolution of latent communities contains the evolution of nodes, the evolution of relationships among nodes, the evolution of community structures, and the changes of structure positions for each community. Since our focus is the evolution of research topics, we adopt six different forms of evolution as suggested in Palla, Barabasi and Vicsek (2007).

The six forms are *Birth*, *Growth*, *Merging*, *Contraction*, *Splitting* and *Death*. Each of them requires analyzing the community structure in both time stamp *t* and time stamp *t+1*. As a result, we simplify the evolution analysis as finding the appropriate successors and predecessors except the *Death* of one community in which there are no successors. Therefore, to detect the relationship between predecessors and successors, we should find predecessors of communities from backward to forward.

Searching the predecessors and successors are essentially the problem of measuring the similarity between two communities. We assume in this paper that if the similarity between two consecutive communities is larger than certain threshold, those two communities have evolution connections. We defined the predecessor of community $M_{(t+1)j}$ as the following formula, in which δ is the threshold value and *d* measures the similarity.

$$Pre(M_{(t+1)j}) = (M_{ti}|M_{ti} \in G_t, d(M_{ti}, M_{(t+1)j}) < \delta) \cup \underset{M_{ti} \in G_t}{\operatorname{argmax}}(d(LM_{ti}, LM_{(t+1)j}))$$

The similarity measurement *d* can be either based on the overlap of nodes (Palla et al., 2007) or based on the *structure similarity* (Berger-Wolf & Saia, 2006). In this paper, we extended the former one and proposed to use FS shown in the following Formula as the similarity measure. In this formula, $EJ(N_x, N_y)$ measures the overlap of nodes, $HS(M_x, M_y)$ measures the overlap of core nodes and $ES(M_x, M_y)$ measures the structure similarity. FS considers the similarity from all of the three aspects.

$$FS(M_x, M_y) = EJ(N_x, N_y) * HS(M_x, M_y) * ES(M_x, M_y)$$

2.4 Community Evolution Visualization

Rosvall and Berstrom (2010) learned the idea of alluvial diagram from Geographic domain to visualize the evolution of networks. Figure 2 provided an example of alluvial diagram. In this figure, the colored rectangle areas represent each community; the colored curve areas between two time stamps denote the evolution process: if one colored rectangle area in time stamp *t* divides into two same colored areas in time stamp *t+1*, it implies that one community divides into two communities; if two colored rectangle areas in time stamp *t* merges into the same colored area in time stamp *t+1*, it implies that two communities merge into one large community, or a new community is created.

Learning from the ideas of alluvial diagram, we develop a colored network diagram (see Figure 3) in visualizing the networks. This diagram is able to represent more details of the community evolution process by providing the successors and predecessors for each node.

There are two types of coloring algorithms in our software: the *backward coloring* algorithm and the *forward coloring* algorithm. The backward coloring helps uncover the future change of the nodes in each community and the forward

coloring helps reveal the source of current community. For example, in the time stamp *Time 1*, Figure 3 has one community A, but in the time stamp *Time 2*, we have two communities B and C who are the successors of community A. The backward algorithm will assign different colors for the nodes in community A based on their community divisions in time stamp *Time 2*. Assumed that in the future in the time stamp *Time 3*, community B and community C merges into one large community D. The forward coloring algorithm will assign different colors for nodes in community D based on their community divisions in time stamp *Time 2*.

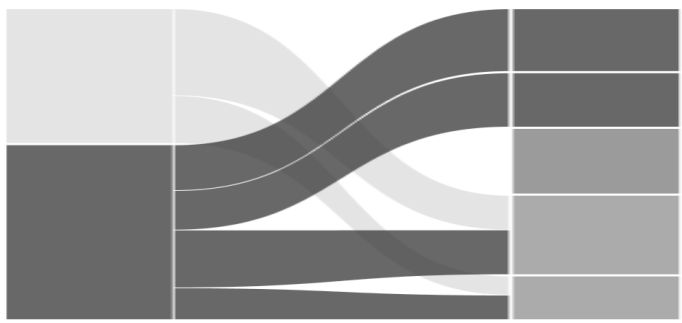


Figure 2. Example of the alluvial diagram

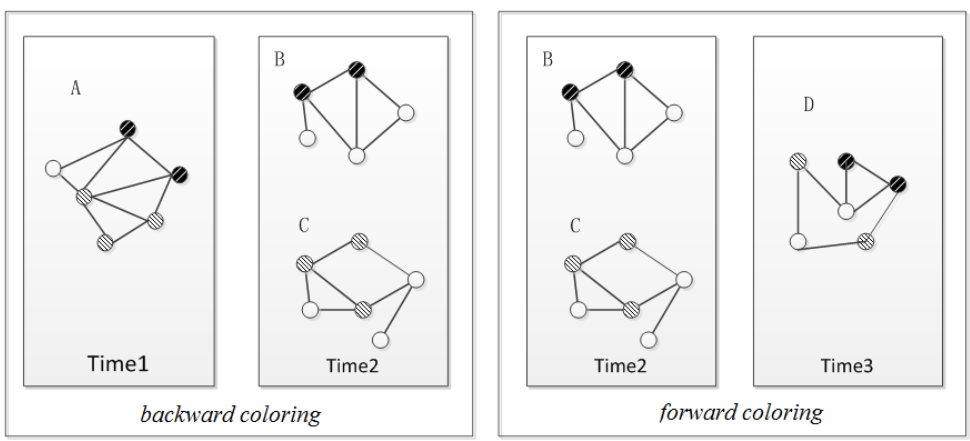


Figure 3. Backboard Coloring and Forward Coloring

We formalize the rules of backward coloring and forward coloring as follows: i) *Backward Coloring* algorithm: given the community M_t in time stamp t, for any node (keyword) v in this community, if the same keyword v also occurred in community $cM_{t+1,i}$, we let $VColor(v) = AColor(cM_{t+1,i})$, in which $cM_{t+1,i}$ represents the ith successor community of community M_t. AColor(M) represents the color of community M_t in the alluvial diagram, VColor(v) denotes the color of node v in

community M_t. ii) *Forward Coloring* algorithm: given the community M_{t+1} in time stamp t+1, for any node (keyword) *v* in this community, if the same keyword *v* also occurred in community $pM_{t,\ i}$, we let $VColor(v) = AColor(pM_{t,\ i})$, in which $pM_{t,\ i}$ represents the ith predecessors community of community M_t.
Besides, we adopt the hierarchical layout in the colored network and put the core nodes in the centre of the layout, because we need to emphasise the importance of the roles of each node in the evolution process.

3. The NEViewer

Based on the previous analysis, we designed and developed a novel network evolution analysis and visualization software: the NEViewer (Network Evolution Viewer). We implement all the algorithms in Java, and we supported the NWB file format (Network Workbench File format)[121]. The developers can also build their own plugins and implement their own algorithms in our software.
NEViewer implement all of the algorithms we mentioned above. For community detection, NEViewer supports Blondel algorithm (Blondel et al., 2008), Newman's Modularity Maximization algorithm (Newman, 2004), Ball's algorithm on overlap community detection (Ball, Karrer, & Newman, 2011). NEViewer also supports different types of community similarity measures such as Jaccard similarity, Tanimoto similarity, cosine similarity, overlap of core nodes and FS measure we mentioned in this paper. NEViewer can visualize the evolution in both the alluvial diagram and the colored network diagram. Users can even choose the network layout. Some basic metrics are also supported in this software such as the PageRank value and the centrality measures.

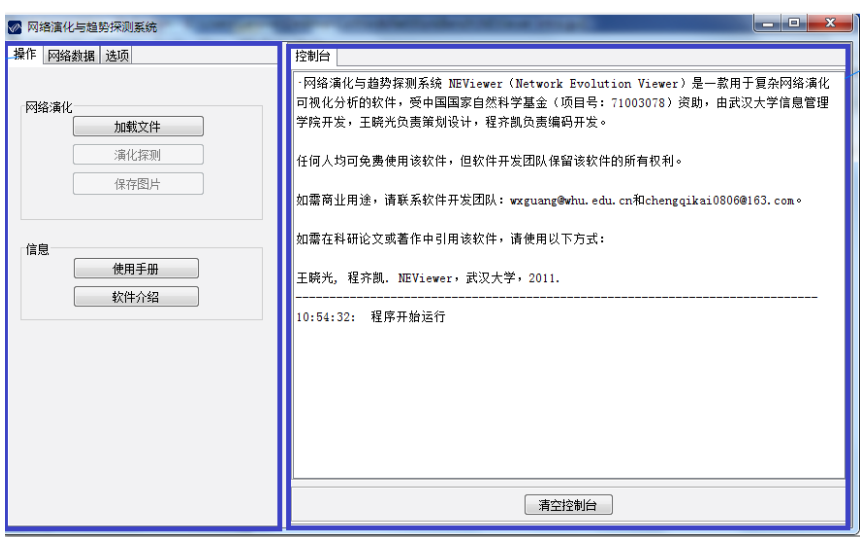


Figure 4. Main view of NEViewer

[121] http://nwb.cns.iu.edu/

4. NEViewer Case Study

4.1 Dataset

In order to evaluate the effectiveness of NEViewer, we constructed an evaluation dataset and conducted a qualitative evaluation on the software. A FIVE-CONF dataset was collected, which contains five main conferences in Information Retrieval, Data Mining and World Wide Web (i.e. KDD, SIGIR, CIKM, CSCW and JCDL). The FIVE-CONF dataset contains 7,234 papers that were published between 2000 and 2011 in one of the five conferences. We exclude workshop papers. Each paper in the dataset includes a title and an abstract. Both stemming and stop word removing were applied on the data.

4.2 Construct Sequential data

The dataset is divided into three parts according to publishing time for constructing the sequential datasets: T1= [t_{2000}, t_{2003}], T2= [t_{2004}, t_{2007}] and T3= [t_{2008}, t_{2011}]. There are 2480 papers in T1, 4283 papers in T2 and 5517 papers in T3. We construct three keyword networks N1, N2 and N3 based on those papers. In total, 1217 nodes (keywords) and 12076 links exist in N1, 2295 nodes (keywords) and 25678 links exist in N2 and 2903 nodes (keywords) and 33272 links exist in N3. The basic network measurements are shown in Table 1. After constructing the sequential data, the Blondel's community detection approach is adopted (Blondel, Guillaume, Lambiotte, & Lefebvre, 2008) and the community detection results are shown in Table 2.

Table 1. Basic Network measures for three datasets

Measurements	*Datasets*		
	T1	T2	T3
Nodes	1217	2295	2903
Isolated Nodes	9	16	17
Edges	12076	25678	33272
Mean degree	19.846	22.377	22.922
largest connected component	1202	2263	2850
Density	0.00816	0.00488	0.00395

Table 2. Number of communities and the average number of nodes in each community in different datasets

Datasets	Number of Communities	Average number of nodes in each community
T1	23	50.71
T2	34	65.57
T3	41	69.12

4.3 Topic Evolution

To visualize the overall keyword networks evolution, we adopt the alluvial diagram proposed by Rosvall and Berstrom (2010) . Figure 5 gives an overall picture of the topic evolution of five conferences from 2000 to 2011, in which we ignore the communities with 10 nodes or less because they usually have little influence on the whole datasets and only present too much noise. The overall evolution provides insights of the evolution of different topics. The major topics involve Information Retrieval (IR), Computer Supported Cooperative Work (CSCW), Social Networks, Visualization, World Wide Web and etc. Those topics are within our expectation because the chosen five conferences are majorly concerning on those five domains. The evolution diagram actually show different types of evolution forms as mentioned in previous study (Palla, Barabasi & Vicsek ,2007). The "information retrieval" community in 2004-2007 divides into several small communities in 2008-2011. The "interaction design" community and partially "education" community in 2000-2003 merged in 2004-2007 into a big "interaction design" community.

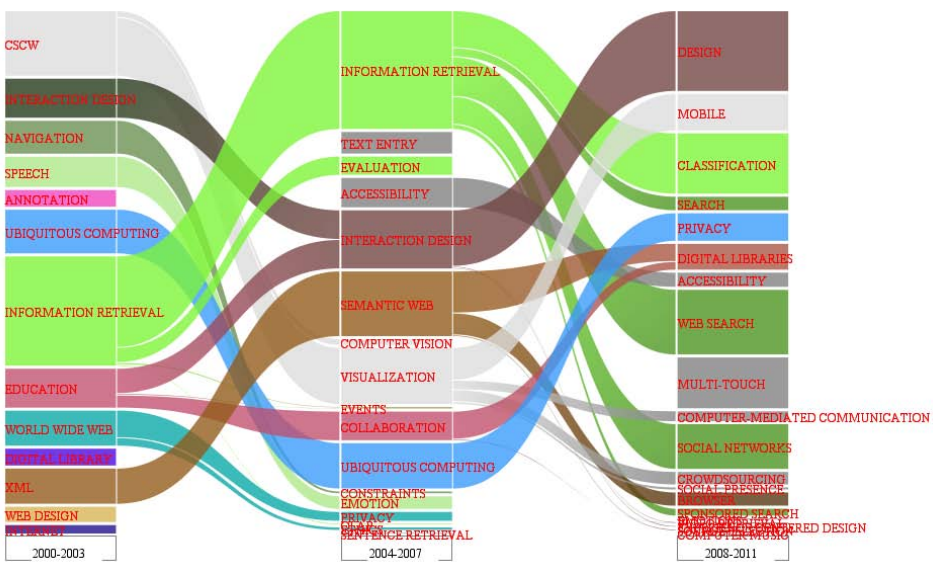


Figure 5. The overall topic evolution on FIVE-CONF dataset

In order to look into the details of how each topic evolves, we took the information retrieval domain as an example. By clicking the Information Retrieval community in our software, we can acquire Figure 6, in which only we only focus on one community evolution of Information Retrieval from T2 to T3. In T3, there seemed to have five different topics: the machine learning application of Information Retrieval (labeled "classification"), the basic Search community,

the Web Search community, the Social Networks analysis in Information retrieval and the Sponsored search community.

In order to evaluate whether the community detection results actually revealed the true latent structures of conference topics. We manually collect the program sessions SIGIR conference from 2008-2011. The results were shown in Table 3. We found that during 2008 and 2010, there was at least a session about ***classification***, which can be used to explain our detected community label "classification". Similarly, almost every year, there was a session named "**Web IR**" or "**Web Search**", which might be corresponding to our detected community labeled "Web IR". Social media and Web 2.0 became important in recent years, and almost every year, there was a session about the "**Social Media**". We also found a small community in our detected community named "sponsored search", there was actually one session in 2010 "***Link Analysis & Advertising***". For all of those detected communities, we actually found similar conference sessions in the real conferences, which actually demonstrated the effectiveness of our community detection and evolution algorithms.

We further analyze the forward colored network diagram of those five communities in Figure 7. The green nodes are those who originate from the previous community. The size of nodes reflects the frequency of each node occurred in the network. In Figure 7(a), we found the "classification" is only a label of this community. The corresponding community actually contains machine learning and data mining technologies. Similarly, the Figure 7(e) showed that the "sponsored search" community actually stands for the computation advertisement; the commonly used algorithms are from link analysis and PageRank.

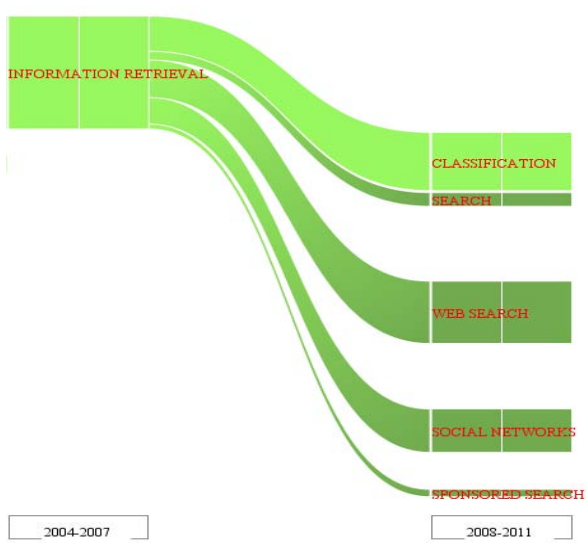
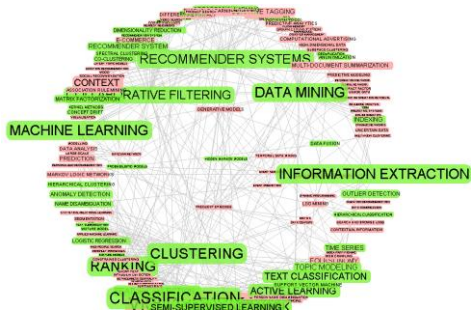
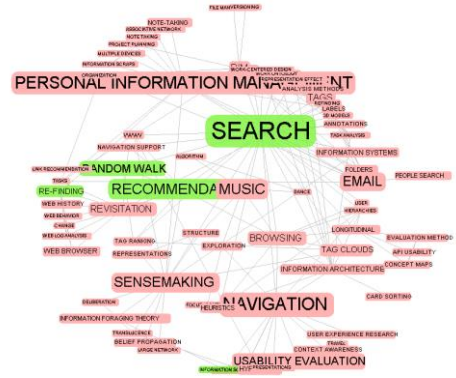


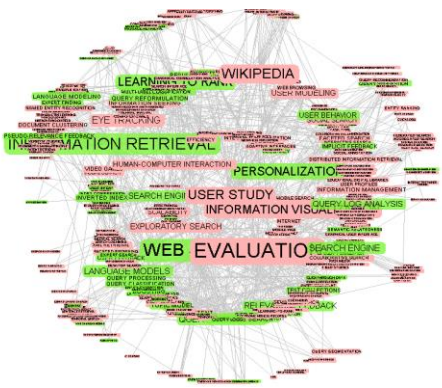
Figure 6: Topic evolution of Information Retrieval from T2 to T3



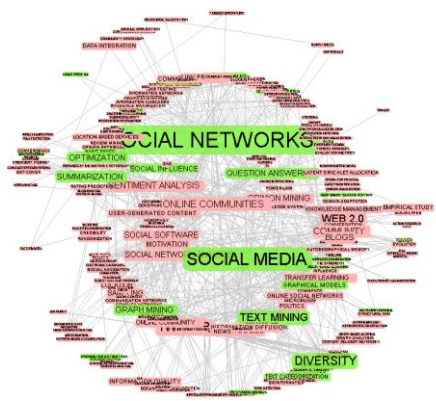
(a) backward colored network for "Classification"



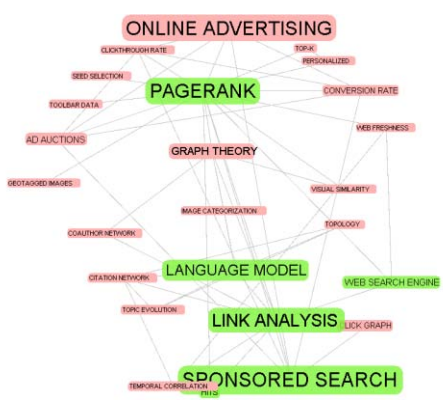
(b) backward colored network for "Search"



(c) backward colored network for "Classification"



(d) backward colored network for "Search"



(e) backward colored network for "Search"

Figure 7: Backward colored network diagram for five communities

Table 3. SIGIR Conference sessions from 2008-2011

2008	*2009*	*2010*	*2011*
User Interaction Models	Novel search features	Clustering	Query Analysis
Web Search	***Classification and clustering***	User Model	Learning to Rank
Evaluation	Expansion and feedback	Applications	Retrieval models
Collaborative Filtering	**Web 2.0**	Search Engine Architectures and Scalability	**Social Media**
Learning to Rank	Retrieval models	***Link Analysis & Advertising***	**Web IR**
High-Performance & High Dimensional Indexing	Speech and linguistic processing	Learning to Rank	Collaborative filtering
User Adaptation & Personalization	Recommenders	Filtering and Recommendation	Query Analysis
Clustering	Question answering	Information Retrieval Theory	Communities
Multilingual & Crosslingual Retrieval	Efficiency	Language Models & IR Theory	Image Search
Relevance Feedback	**Web retrieval**	Query Representations & Reformulations	Web Queries
Summarization	Learning to Rank	***Automatic Classification***	Collaborative filtering
Exploratory Search & Filtering	Information extraction	Retrieval Models and Ranking	Multimedia IR
Multimedia Retrieval	Clickthrough models	User Feedback & User Models	Summarization
Query Analysis & Models	Vertical search	**Web IR and Social Media Search**	Query suggestions
Non-Topicality	Interactive search	Document Structure & Adversarial Information Retrieval	Linguistic Analysis
Probabilistic Models	Multimedia	Users and Interactive IR	Effectiveness
Analysis of Social Networks	Federated, distributed search	Document Representation and Content Analysis	Multilingual IR
Question-Answering	Industry Track speakers	Test-Collections	Recommender systems
Social Tagging	Evaluation and measurement	Query Log Analysis	Test collections
Content Analysis	Query formulation	Summarization & User Feedback	
Learning Models for IR	Spamming	Query Analysis	
Text Classification		Effectiveness Measures	
		Multimedia Information Retrieval	
		Non-English IR & Evaluation	

5. Conclusion and Discussion

In order to analyze the evolution of scientific topics, we proposed a novel co-keyword network evolution based method and developed the software, i.e. the NEViewer to tackle with the community detection, community mapping, representative node finding and evolution visualization. The NEViewer provides implementations of several well-known community analysis algorithms so that users can specify their own preferences.

The NEViewer consists of four steps in the real analysis: the topic(community) finding, the representative node selection, the evolution detection and the visualization. We developed each step in different modules so that the developers can implemented their own algorithms to each of our four steps.

There are already several topic evolution and knowledge mapping software such as CiteSpace, VosViewer, Network Workbench and SciMat. However, comparing to those software, the novelty of NEViewer are: (a) we developed four independent steps in analyzing network visualization; (b) we implemented the alluvia diagram in order to show the visualization and we also designed and

implemented the Backward and Forward colored network diagram in helping users make sense of both the macro- and micro- level of network evolution.
However, there are still some limitations on our current approaches. We didn't distinguish the functional roles of each keyword. Some keywords may represent the methodology while others reflect the research objects. This makes it difficult for us to find the accurate representative nodes in each community. Besides, the community mapping is still a big issue because they lacked a firm threshold and principles of predefining the thresholds.

Acknowledgement

This paper is supported by the National Natural Science Foundation of China (Grant No.71003078) and the National Natural Science Foundation of China (Grant No. 71173164).

References:

Ball, B., Karrer, B., & Newman, M. (2011). Efficient and principled method for detecting communities in networks. *Physical Review E, 84*(3), 36103.

Berger-Wolf, T. Y., & Saia, J. (2006). *A framework for analysis of dynamic social networks*.

Blondel, V. D., Guillaume, J. L., Lambiotte, R., & Lefebvre, E. (2008). Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment, 2008*(10), P10008.

Börner, K., Chen, C., & Boyack, K. W. (2005). Visualizing knowledge domains. *Annual review of information science and technology, 37*(1), 179-255.

Boyack, K. W., Klavans, R., & Börner, K. (2005). Mapping the backbone of science. *Scientometrics, 64*(3), 351-374.

Chen, C. (2005). CiteSpace II: Detecting and visualizing emerging trends and transient patterns in scientific literature. *Journal of the American Society for Information Science and Technology, 57*(3), 359-377.

Cobo, M. J., López-Herrera, A. G., Herrera-Viedma, E., & Herrera, F. (2010). An approach for detecting, quantifying, and visualizing the evolution of a research field: A practical application to the fuzzy sets theory field. *Journal of Informetrics*.

Ding, Y. (2011). Community detection: Topological vs. topical. *Journal of Informetrics, 5*(4), 498-514.

Guimerà, R., Sales-Pardo, M., & Amaral, L. A. N. (2006). Classes of complex networks defined by role-to-role connectivity profiles. *Nature physics, 3*(1), 63-69.

Lancichinetti, A., & Fortunato, S. (2009). Community detection algorithms: A comparative analysis. *Physical Review E, 80*(5), 56117.

Leydesdorff, L., & Rafols, I. (2008). A global map of science based on the ISI subject categories. *Journal of the American Society for Information Science and Technology, 60*(2), 348-362.

Mane, K. K., & Börner, K. (2004). Mapping topics and topic bursts in PNAS. *Proceedings of the National Academy of Sciences of the United States of America, 101*(Suppl 1), 5287-5290.

McCain, K. W. (2008). Assessing an author's influence using time series historiographic mapping: The oeuvre of Conrad Hal Waddington (1905–1975). *Journal of the American Society for Information Science and Technology, 59*(4), 510-525.

Newman, M. E. J. (2004). Fast algorithm for detecting community structure in networks. *Physical Review E, 69*(6), 66133.

Palla, G., Barabasi, A. L., & Vicsek, T. (2007). Quantifying social group evolution. *Nature, 446*(7136), 664-667.

Palla, G., Derényi, I., Farkas, I., & Vicsek, T. (2005). Uncovering the overlapping community structure of complex networks in nature and society. *Nature, 435*(7043), 814-818.

Rosvall, M., & Bergstrom, C. T. (2010). Mapping change in large networks. *PloS one, 5*(1), e8694.

Wallace, M. L., Gingras, Y., & Duhon, R. (2009). A new approach for detecting scientific specialties from raw cocitation networks. *Journal of the American Society for Information Science and Technology, 60*(2), 240-246.

Wang, X., Jiang, T., & Li, X. (2010). Structures and dynamics of scientific knowledge networks: An empirical analysis based on a co-word network. *Chinese Journal of Library and Information Science*(3), 19-36.

Lin, X., Zhang, M., Zhao, H., and Buzydlowski, J. 2012. Multi-view of the ACM classification system. *Proceedings of the 12th ACM/IEEE-CS joint conference on Digital Libraries*. 397-398.

THE NUANCED NATURE OF E-PRINT USE: A CASE STUDY OF ARXIV

Vincent Larivière[1], Benoit Macaluso[1], Cassidy R. Sugimoto[2], Staša Milojević[2], Blaise Cronin[2], and Mike Thelwall[3]

[1] *vincent.lariviere@umontreal.ca; macaluso.benoit@uqam.ca*
École de bibliothéconomie et des sciences de l'information, Université de Montréal, C.P. 6128, Succ. Centre-Ville, Montréal, QC. H3C 3J7 (Canada) and
Observatoire des sciences et des technologies (OST), Centre interuniversitaire de recherche sur la science et la technologie (CIRST), Université du Québec à Montréal, CP 8888, Succ. Centre-Ville, Montréal, QC. H3C 3P8, (Canada)

[2] *sugimoto@indiana.edu; smilojev@indiana.edu; bcronin@indiana.edu*
School of Information and Library Science, Indiana University Bloomington
1320 E. 10th St. Bloomington, IN 47401 (USA)

[3] *m.thelwall@wlv.ac.uk*
School of Technology, University of Wolverhampton, Wulfruna Street, Wolverhampton WV1 1LY (UK).

Abstract
Since its creation in 1991, arXiv has become central to the diffusion of research in a number of fields. Combining data from the entire arXiv and the Web of Science (WoS), this paper investigates (a) the proportion of papers across all disciplines that are on arXiv, (b) Elapsed time between arXiv submission and journal publication, and (c) the aging characteristics and scientific impact of arXiv e-prints and their published version. It shows that the proportion of WoS papers found on arXiv varies across the specialties of Physics and Mathematics, and that only a few specialties make extensive use of the repository. Elapsed time between arXiv submission and journal publication has also shortened, but remains longer in Mathematics than in Physics. In Astronomy and Astrophysics, arXiv versions are cited more promptly and decay faster than WoS papers. Unsurprisingly, arXiv versions of papers—both published and unpublished—have lower citation rates, although there is almost no difference in the impact of the arXiv versions of both published and unpublished papers.

Conference Topic
Open Access and Scientometrics (Topic 10) and Bibliometrics in Library and Information Science (Topic 14).

Introduction
Preprints—"temporary documents whose function is to bridge the time-gap created by publication delays" (Goldschmidt-Clermont, 1965, p. 8)—are a well-established mechanism for the exchange of scientific information (Mikhailov,

Chernyi, & Giliarevskii, 1984). This is particularly true in astronomy and physics, disciplines that have long used preprints to communicate research results (Brooks, 2009; Brown, 2001; Wilson, 1970; Kling, 2005) and establish priority claims, thereby effectively reducing the role of the journal to "secondary distribution, archiving, and peer review" (Brooks, 2009, p. 92). Advocates of open access view subject repositories, such as arXiv, as heralding the eventual demise of the scholarly journal and have outlined ways in which peer review might function on these new platforms (Rodriguez, Bollen, & Van de Sompel, 2006) while others look forward to "the stranglehold journal publishers have over science libraries" being broken (Carriveau, 2008, p. 73). Hence the question: can or need these two forms of scholarly communication co-exist (Morris, 2003)?

First a word of caution: one should not be blinded by enthusiasm for the new. Preprints, after all, are far from novel. By way of illustration, the Information Exchange groups, run by the National Institutes of Health, circulated more than 1.5 million preprints in 1966 (Confrey, 1996). Moreover, relatively few scholars, with physicists and mathematicians being notable exceptions, use preprints extensively (Swan & Brown, 2003). Lastly, what appears to work as a publishing model in one field may not translate to another (Kling, Spector, & McKim, 2002; Kling, Spector, & Fortuna, 2004).

Since its creation by Paul Ginsparg in 1991, arXiv has become central to the diffusion of research in a number of related fields: physics, mathematics, and computer science in particular (Gentil-Ceccot, Mele, & Brooks, 2009). Previous research has examined: the use made of arXiv (Brown, 2001); ordering and citation rates (Haque & Ginsparg, 2009); the coexistence of e-prints and journals (Henneken et al., 2007); and the effect of arXiv on citation rates (Moed, 2007). However, data from all of arXiv and the Web of Science (WoS) have yet to be combined for a comparative analysis. This paper combines the entire arXiv with the entire Web of Science in order to better understand the ecology of scholarly communication. More specifically, we investigate (a) the proportion of papers across all disciplines that are on arXiv, (b) the elapsed time between arXiv submission and journal publication, and (c) the aging characteristics and scientific impact of arXiv e-prints and their alter egos (the versions published in the journal of record). This last analysis is performed on a subset of the dataset comprising papers published in Astronomy and astrophysics.

Background

arXiv and related platforms

Paul Ginsparg launched xxx.lanl.gov, the first Internet-based e-print server, in 1991 to facilitate preprint exchange in the field of theoretical high-energy physics (Brown, 2010; Carriveau, 2008; Davis & Fromert, 2007; Ginsparg, 2008). The name was changed to arXiv.org in 1998, after it grew in popularity and expanded

to cover other fields (Ginsparg, 2008). The aim was to create an electronic bulletin board "to serve a few hundred friends and colleagues" (Ginsparg, 2011, p. 145). arXiv was "designed as a way of automating a paper-based process already in existence" (Pre-print culture, Pinfield, 2001, para. 1). Today's much-enlarged arXiv is strongest in physics, mathematics, and computer science (Brody, Harnad, & Carr, 2006), fields in which there is a tradition of preprint use.

The number of articles in arXiv has been growing linearly since 1991 (Brody, Harnad, & Carr, 2006) and arXiv is now the "largest self-archived centralized e-print archive" (Brody, Harnad, & Carr, 2006, p. 102). Originally hosted at the Los Alamos National Laboratory (hence the initial domain name), it was later moved to Cornell University, where it is under the aegis of the university library (Hey & Hey, 2006). In parallel, Ginsparg began a movement to develop a set of technical standards for the establishment of a global preprint archive via the Universal Preprint Service Initiative—later known as the Open Archives Initiative (Brown, 2001; Manuel, 2001). The federated nature of OAI repositories has led to proposals for a "repository-centric peer-review model" based on the OAI platform and using a social-network algorithm to suggest potential reviewers and weigh evaluations (Rodriguez, Bollen, & Van de Sompel, 2006).

In 1997, arXiv began collaborating with the Astrophysics Data System (ADS). The ADS created an index for the astrophysics e-prints and made them available through the ADS abstracts service. In 2002, abstracts of all arXiv categories were included (Henneken et al., 2007). arXiv also has a relationship with SPIRES, the first electronic catalogue of grey literature, focused on high-energy physics preprints (Bentil-Beccot, Mele, Brooks, 2009). SPIRES counts citations to and from preprints and directs physicists to arXiv (82% of clicks from SPIRES go to arXiv) (Bentil-Beccot, Mele, Brooks, 2009). SPIRES is currently being replaced with INSPIRE, which was created to "provide an even more flexible and extensible system to allow publishers, repositories, and researchers themselves to contribute and share information" (Brooks, 2009, p. 91). A survey of high-energy physicists found that nearly 90% rely on SPIRES and arXiv as their point of entry to the literature. This system is so embedded in the working practice of physicists that Kling, McKim, and King (2003) considered SPIRES, arXiv, and associated human actors as the embodiment of a functioning socio-technical interaction network.

Empirical investigations of arXiv

Over the years, several studies have focused on authors' practices with respect to arXiv: Fowler's (2011) survey of mathematicians found that 81% had posted to arXiv and that it was a regular sharing mechanism for 30%; Manuel (2001) found that authors were primarily academic (rather than corporate); and Moed (2007) showed that posters tended to be high-impact authors (measured by the citation impact of those of their papers not deposited in arXiv). However, most research

has focused on the preprints—specifically on the relationship between preprints and their subsequent publication and impact. For example, approximately half of all preprints in arXiv are subsequently published in peer-reviewed publications (Manuel, 2001; Mine, 2009). Studies have also looked at the inverse, viz., the proportion of journal literature in a given field that is also in arXiv. Rates included almost 100% in high energy physics (Gentil-Beccot, Mele, & Brooks, 2009), 75% in condensed matter (Moed, 2007), and 18.5% in mathematics (Davis & Fromerth, 2007). The number of articles appearing in both arXiv and the published literature is increasing (Gentil-Beccot, Mele, & Brooks, 2009; Davis & Fromerth, 2007). Peer-reviewed articles that were also preprints receive significantly more citations than articles not deposited (Davis & Fromerth, 2007; Gentil-Beccot, Mele, & Brooks, 2009). The reasons suggested are: an early view effect, a quality differential, and an open access advantage (Kurtz et al., 2005; Davis & Fromerth, 2007).

Some studies confirm the early view effect (Moed, 2007): "colleagues in the field start the process of reading a paper, processing its information, and citing it in their own articles earlier if a paper is deposited in arXiv" (Moed, 2007, p. 2053). However, other studies have found no such effect (Davis & Fromerth, 2007). Evidence has also been found to support a 'quality bias', that is, better papers and high impact authors appear in arXiv more than the reverse (Davis & Fromerth, 2007; Moed, 2007). Little or no support has been found for the open access advantage, however (Moed, 2007; Kurtz et al., 2005; Davis & Fromerth, 2007). As Kurtz et al. (2005, p. 1400-1401) concluded: "This implies that there is no significant population of astronomers who are both authors of major journal articles and who do not have 'sufficient' access to the core research literature." Haque and Ginsparg (2009, 2010) found that posts on arXiv at the beginning and end of the day receive higher levels of citation and readership than those in the middle. Other studies have examined the proportion of citations to the e-print version of the paper, with mixed findings (Youngen, 1998; Manuel, 2001).

Readership has been investigated, too. Using two years of cumulative download and citation data from arXiv, Brody, Harnad and Carr (2006) found that download counts at six months provided reliable predictions of citation impact at two years. They concluded that, "the rapid dissemination model of arXiv has accelerated the read-cite-read cycle substantially" (Harnad, Brody, & Carr, p. 1062). The relationship between the publisher's version and the preprint remains unclear: Davis and Fromerth (2007) found that arXiv-deposited articles received 23% fewer downloads from publishers' websites. However, Henneken et al. (2007), in a study of four astronomy journals, found that reads of the arXiv e-print through ADS dropped to zero (or near zero) immediately following the publication of the peer-reviewed article. They also note that the half-life of e-prints is shorter than that of the corresponding journal articles, concluding that, "e-prints have not

undermined journal use in the astrophysics community and thus do not pose a threat to the journal readership" (Henneken et al., 2007, p. 19).

Methods

Here we use two data sources: the arXiv database and WoS. All arXiv database metadata from 1990 to March 22, 2012 were downloaded (N = 744,583 e-prints). All standard citation indexes were used for WoS (Science Citation Index Expanded, Social Sciences Citation Index and Arts and Humanities Citation Index) for the 1990-2011 period. Data are presented for 1995—2011 (although citations and matching papers were compiled until the first 42 weeks of 2012). Two types of links between the data sources were created: (a) between the arXiv e-print and its published version indexed in WoS and (b) between the arXiv e-print and the citations it received in WoS. Several steps were needed to match the arXiv e-print to its published counterpart (a). First, three sets of links were established: 1) direct correspondence between the arXiv and WoS titles, 2) fuzzy matching between the arXiv and WoS titles AND fuzzy matching between the journal mentioned in the arXiv bibliographical notice and the WoS journal, 3) fuzzy matching between the arXiv and WoS titles AND fuzzy matching between the arXiv first author and the WoS paper first author. These links were, in a second step, automatically validated through the similarity of their abstracts. In total, 441,018 out of the 744,583 arXiv e-prints (59.2%) were matched with a WoS-indexed journal article, note or review.

For the second matching (b) we utilized the specific structure of the references to the arXiv e-prints in WoS. For example, a reference to an e-print from the condensed matter section of arXiv will have the string 'CONDMAT' followed by the series of seven or eight digits that correspond to its document ID in the online e-print database. Given that a paper belonging to more than one arXiv category can be cited using both categories as prefixes, the matching process used the seven or eight digits as well as its prefix. For Astronomy and Astrophysics, we separated documents into four distinct categories: 1) arXiv e-prints never published in a WoS-indexed journal, 2) arXiv e-prints published in a WoS-indexed journal, 3) WoS-indexed journal articles also published and archived as an arXiv e-print, and 4) WoS-indexed journal articles that were never published as arXiv e-prints. Finally, the field classification used is that of the U.S. National Science Foundation[122], developed by *The Patent Board*.

Results and discussion

Proportion of WoS papers on arXiv

As mentioned in the Methods section, about 60% of all arXiv e-prints are published in a WoS-indexed journal. This percentage is slightly higher than those

[122] http://www.nsf.gov/statistics/seind06/c5/c5s3.htm#sb1

obtained by Manuel (2001) and Mine (2009), which is likely a consequence of the increase in self-archiving in recent years. On the other hand, when taking all WoS papers as the denominator, only 3.3% of 2010-2011 WoS papers (all disciplines combined) were submitted to arXiv. Three disciplines account for the vast majority (93%) of arXiv submissions in 2010-2011: Mathematics (with 21% of all WoS papers on arXiv), Physics (19% of all WoS papers on arXiv) and Earth and Space (11% of all WoS papers on arXiv). Within these disciplines, a few specialties are using it more intensively. As shown in Figure 1, about two-thirds of WoS papers published in Astronomy and Astrophysics and Nuclear and Particle Physics are found on arXiv. The Inset of Figure 1 shows that this percentage has increased since 1995. While researchers in Nuclear and Particle Physics were quick to adopt arXiv—this percentage was already greater than 60% in 2000—those in astronomy gradually made a greater use of it. Since the mid-2000s, both specialties have used arXiv to the same extent. In Nuclear and Particle Physics, the percentage we obtain is lower than that of Gentil-Beccot, Mele, and Brooks (2009) for high-energy physics, which is due to the fact that their definition of the field only included 5 high-impact journals, while ours covered 48 journals. In Mathematics, our percentages are higher than those of Davis and Fromerth (2007), which is likely a consequence of the increase of papers appearing in both arXiv and in the WoS.

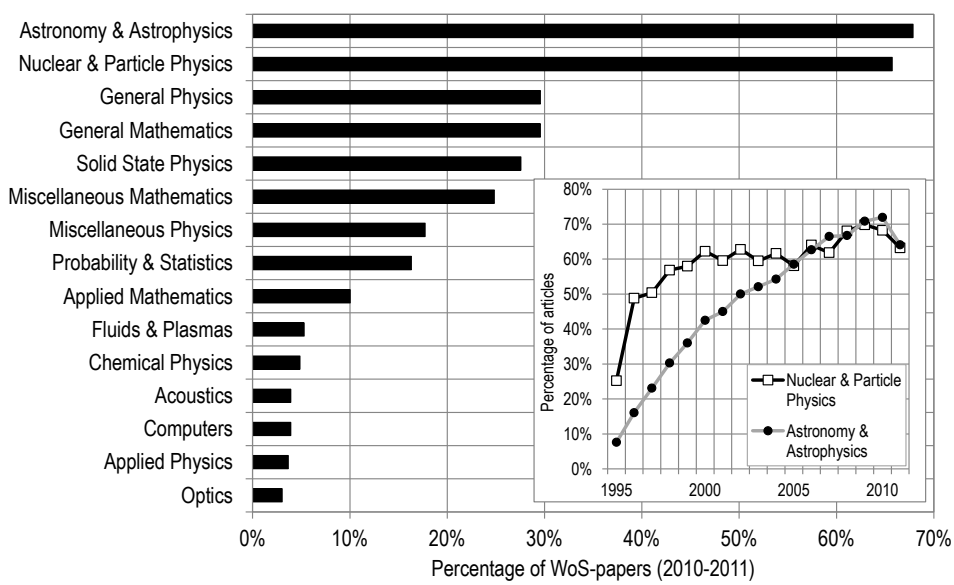


Figure 1. Proportion of WoS papers on arXiv, by specialty (2010-2011). Inset: Proportion of WoS papers on arXiv, by specialty, 1995-2011.

Elapsed time between arXiv submission and journal publication Figure 2 shows that the time between the submission of the manuscript to arXiv and publication

in a peer-reviewed journal has decreased[123]. Whereas papers were once published a year after appearing on arXiv, publication in a journal is now likely to occur in the same year as their appearance on arXiv. There are two possible reasons for this: 1) a higher proportion of researchers are waiting for the paper to be published or accepted for publication before submitting to arXiv, or 2) the introduction of arXiv may have motivated publishers to try to reduce publication delays.

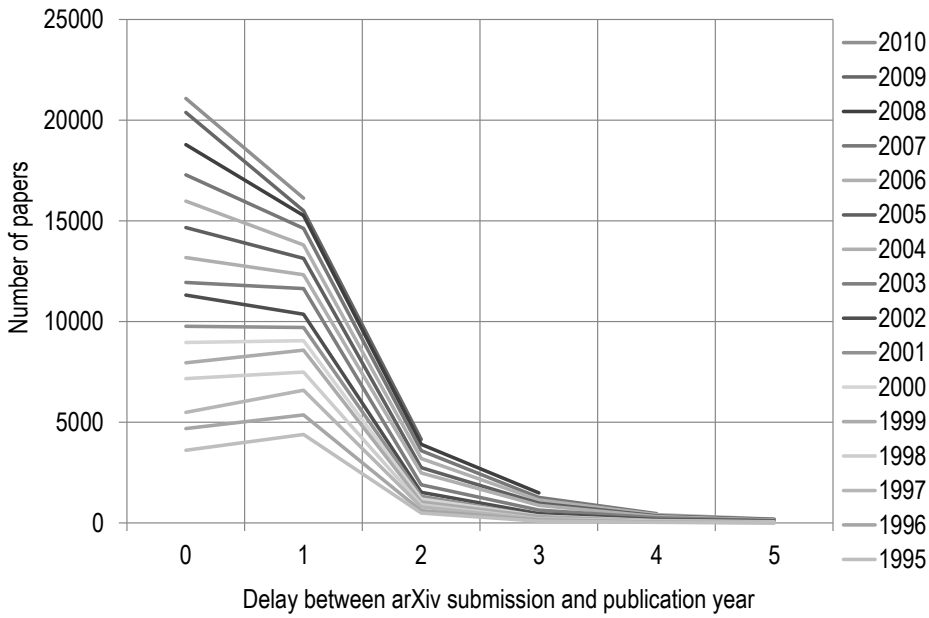


Figure 2. Distribution of the elapsed time between arXiv submission and publication year, by year of submission to arXiv, 1995-2010.

Elapsed time between arXiv submission and journal publication varies dramatically across specialties of science. Figure 3 presents this interval—compiled as an average—for the 18 specialties with more than 1,000 WoS papers found on arXiv. It globally shows that specialties of physics have very short delays—less than half a year on average—while those of mathematics have longer delays (>1 year). Among the specialties with the shortest time between arXiv submission and journal publication is Astronomy and Astrophysics, one of the two specialties with the most intensive use of preprints. The appearance of General Biomedical Research is due to the fact that "general" journals that publish Physics or Mathematics papers, such as *Science* and *Nature,* are categorized in this specialty.

[123] 11,946 e-prints out of 440,371 that matched to a WoS paper (2.7%) have been submitted on arXiv after journal publication; those have been removed from this part of the analysis.

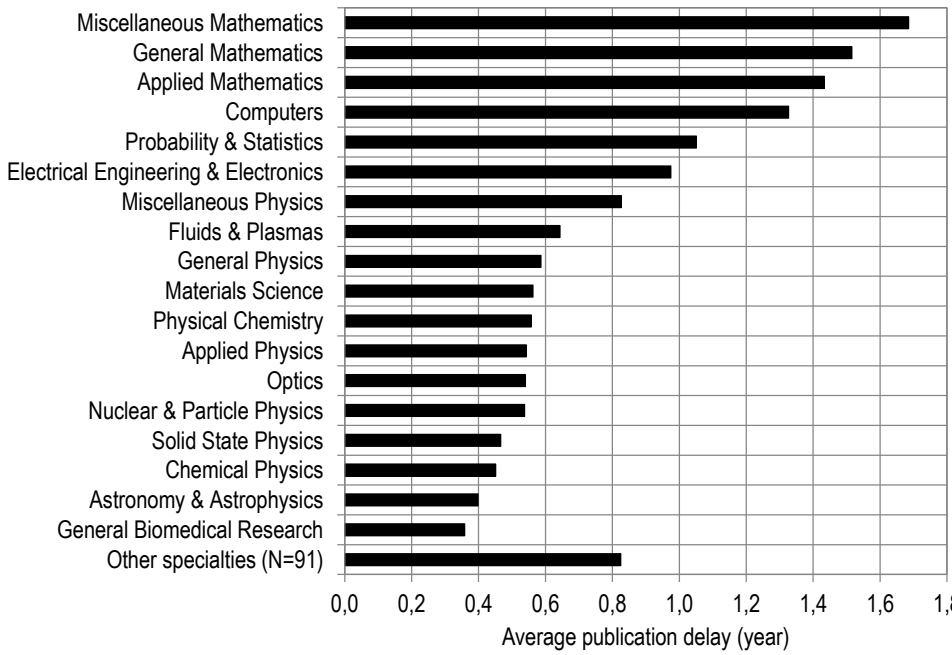


Figure 3. Average elapsed time between arXiv submission and journal publication, by specialty (with more than 1,000 e-prints matched with a WoS paper), 1995-2011.

Aging characteristics and scientific impact

This section analyzes the aging characteristics and scientific impact of arXiv e-prints and their WoS-published alter egos for the specialty Astronomy and Astrophysics. Figure 4 presents (A) the trends in the numbers of papers that have appeared on arXiv only, on arXiv and WoS (arXiv version), in WoS only and on arXiv and WoS (WoS version) and (B) the mean number of citations these documents have received using a one-year citation window plus publication year. We see a considerable increase in the number of documents published both in arXiv and in journals, a small increase in the number of papers published only in arXiv, and a decline in papers published only in journals. In terms of proportion of all distinct Astronomy and Astrophysics documents in 2010—obtained by the combination of both arXiv and WoS—19% are found on arXiv only, 54% both on arXiv and in WoS and 27% in WoS only.

The citation rates of the four groups are quite different and vary over time. WoS versions of arXiv e-prints obtain the highest citation rates, a finding consistent with the well documented association between arXiv submission and citation (Davis & Fromerth, 2007; Gentil-Beccot, Mele, & Brooks, 2009). However, this mean impact is decreasing—even when we add to the WoS version the citations received by the arXiv version—and is approaching that of other WoS papers not submitted to arXiv, whose mean impact is increasing. arXiv versions—both

published and unpublished—obtain lower citation rates. Surprisingly, however, there is almost no difference in the impact of the arXiv versions of both published and unpublished papers. One could have expected that these unpublished papers, being non-refereed, would have a lower impact than comparable arXiv submissions published in a journal. However, it is possible that researchers prefer to cite the published version of an e-print which is likely to reduce published e-print impact and, hence, make the two measures comparable. On the whole, these results are consistent with those of Brooks (2009), who showed that unpublished arXiv submissions had 5 times less impact than those published in a journal, when one includes the citations received by the published version of the e-print.

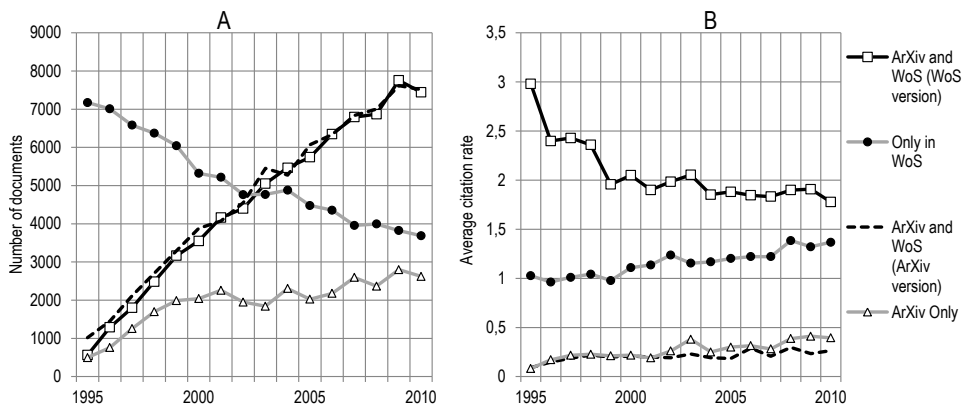


Figure 4. A) Number of documents published and B) mean number of citation received (publication year plus one year), for documents published on arXiv only, on arXiv and WoS (arXiv version), only in WoS and on arXiv and WoS (WoS version), 1995-2010

In terms of aging characteristics, Figure 5 presents the age distribution of citations received by the four groups of documents. It shows that e-prints and published papers follow different patterns. Citations to e-prints peak the year following submission, while citations to papers are similar during the two years following the publication year. Given the transfer of citations from pre-publication e-prints to their published version (Brown, 2001; Henneken et al., 2007), citations to their e-print versions decay faster than those received by unpublished e-prints. E-prints found on arXiv only have a slower decay, although it is faster than that of WoS papers. These faster citations for arXiv e-prints are consistent with findings of Harnad, Brody, and Carr (2006) as well as those of Henneken et al. (2007).

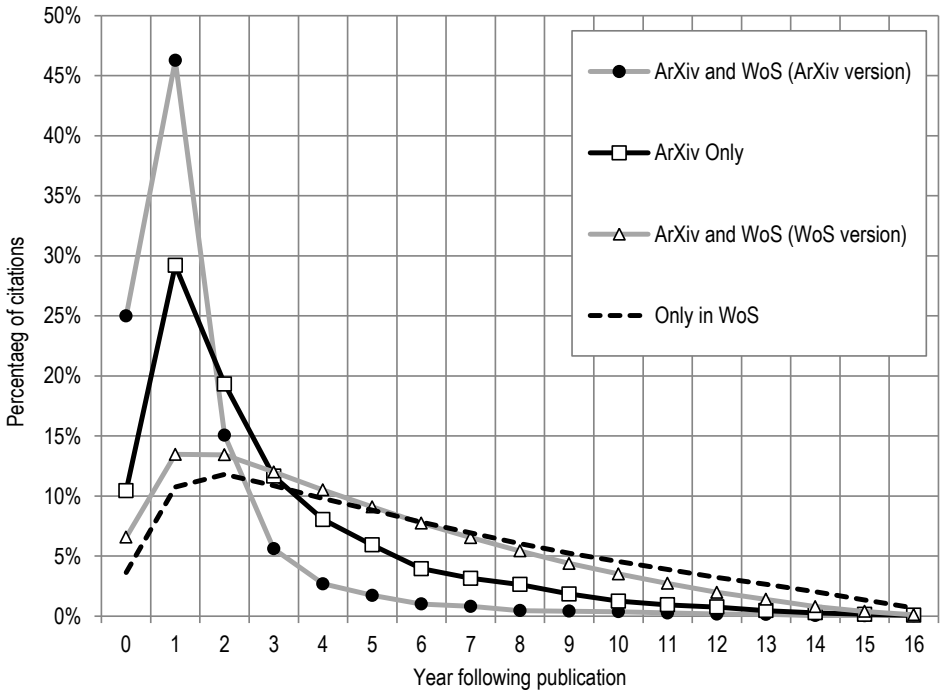


Figure 5. Percentage of citations received, for documents published on arXiv only, on arXiv and WoS (arXiv version), only in WoS and on arXiv and WoS (WoS version), 1995-2010

Conclusion

This paper shows that arXiv has changed the scholarly communication patterns of physicists and mathematicians. In some specialties, such as Astronomy and Astrophysics and Nuclear and Particle Physics, the vast majority of papers published in WoS-indexed journals are found on arXiv. The role of arXiv in these communities has moved from the space of sharing pre-prints by minority, to the place for archiving the majority of produced research. However, we also note that, in those disciplines, there is still a significant proportion of papers that are not on arXiv. Previous research on the topic, focusing on high-impact journals exclusively, has found a greater proportion of WoS-papers in those specialties to be on arXiv (Gentil-Beccot, Mele, & Brooks, 2009), Our results show that, when the whole discipline is considered—both high-impact and low-impact journals alike—the proportion of published papers that are self-archived on arXiv is noticeably lower. Similarly, not all specialties are using it to the same extent: in most specialties of Physics and Mathematics, less than a third of WoS papers are found on arXiv. Along these lines, arXiv is increasingly used outside these two fields, but is still quite marginal: 93% of all WoS-published arXiv e-prints are published either in Mathematics, Physics or Earth and Space sciences. Our results also show that the average elapsed time between submission to arXiv and

publication in a WoS-indexed peer-reviewed journal has decreased over time. This is either due to a higher proportion of researchers waiting for the paper to be published or accepted for publication before submitting to arXiv, or to a reduction in publication delays. These time lags are also quite different across fields of science, with Physics specialties having shorter delays than specialties of Mathematics.

The subset of Astronomy and Astrophysics papers analysed shows that arXiv versions of papers are cited more promptly and decay faster than WoS papers. WoS versions of arXiv e-prints obtain the highest citation rates, but the difference with other WoS papers not submitted to arXiv is decreasing. Unsurprisingly, arXiv versions of papers—both published and unpublished—obtain lower citation rates, although there is almost no difference in the impact of the arXiv versions of both published and unpublished papers. As Harnad, Brody, and Carr (2006) point out, the fact that preprints are cited before publication—and, hence, peer review—as well as the fact that unpublished e-prints are cited raises the question of the function of peer-review in those fields. It seems that citing authors either evaluate papers themselves, often being reviewers, or trust the results presented—which might be a consequence of the few massive collaborations and large-scale scientific infrastructures found in these disciplines.

Acknowledgments

This research was funded by SSHRC (Canada), National Science Foundation (U.S.; grant #1208804) AHRC/ESRC/JISC (U.K.) through the *Digging into Data* funding program.

References

Brody, T., Harnad, S., & Carr, L. (2006). Earlier web usage statistics as predictors of later citation impact. *Journal of the American Society for Information Science & Technology*, 57(8), 1060-1072.

Brooks, T.C. (2009). Organizing a research community with SPIRES: Where repositories, scientists and publishers meet. *Information Services & Use*, *29*, 91-96.

Brown, C. (2001). The e-volution of preprints in the scholarly communication of physicists and astronomers. *Journal of the American Society for Information Science & Technology*, *52*(3), 187-200.

Brown, C. (2010). Communication in the sciences. In B.Cronin (Ed.), *Annual Review of Information Science & Technology* (pp. 287-316). Medford, N.J.: Information Today.

Carriveau, K.L. (2008). A brief history of e-prints and the opportunities they open for science librarians. *Science & Technology Libraries*, *20*(2-3), 73-82.

Confrey, E.A. (1996). The Information Exchange Groups experiment. *Publishing Research Quarterly*. *12*(3), 37-39

Davis, P.M., & Fromerth, M.J. (2007). Does the arXiv lead to higher citations and reduced publisher downloads for mathematics articles? *Scientometrics*, *71*(2), 203-215.

Fowler, K.K. (2011). Mathematicians' views on current publishing issues: A survey of researchers. *Issues in Science and Technology Librarianship*, *67*. Retrieved from: http://www.istl.org/11-fall/refereed4.html

Gentil-Beccot, A., Mele, S., & Brooks, T.C. (2009). Citing and reading behaviours in high-energy physics. How a community stopped worrying about journals and learned to love repositories. Retrieved from: http://arxiv.org/abs/0906.5418

Ginsparg, P. (2008). The global village pioneers. *Learned Publishing*, *21*, 95-100.

Ginsparg, P. (2011). arXiv at 20. *Nature*, 476, 146-147.

Goldschmidt-Clermont, L. (1965). Communication patterns in high-energy physics. Retrieved from: http://testing-library.web.cern.ch/testing-library/Webzine/6/papers/1

Haque, A.-u., & Ginsparg, P. (2009). Positional effects on citation and readership in arXiv. *Journal of the American Society for Information Science & Technology*, *60*(11), 2203-2218.

Haque, A.-u., & Ginsparg, P. (2010). Last but not least: Additional positional effects on citation and readership in arXiv. *Journal of the American Society for Information Science & Technology*, *61*(12), 2381-2388.

Henneken, E.A., Kurtz, M.J., Eichhorn, G., Accomazzi, A., Grant, C.S., Thompson, D., Bohlen, E., Murray, S.S., Ginsparg, P., & Warner, S. (2007). E-prints and journal articles in astronomy: A productive co-existence. *Learned Publishing*, *20*, 16-22.

Hey, T., & Hey, J. (2006). e-Science and its implications for the library community. *Library Hi Tech*, *24*(4), 515-528.

Kling, R. (2005). The Internet and unfrefereed scholarly publishing. In B.Cronin (Ed.), *Annual Review of Information Science & Technology* (pp. 591-631). Medford, N.J.: Information Today.

Kling, R., McKim, G., & King, A. (2003). A bit more to it: Scholarly communication forums as socio-technical interaction networks. *Journal of the American Society for Information & Technology*, *54*(1), 47-67.

Kling, R., Spector, L., & McKim, G. (2002). Locally controlled scholarly publishing via the Internet: The Guild model. *The Journal of Electronic Publishing*, *8*(1). doi: http://dx.doi.org/10.3998/3336451.0008.101

Kling, R., Spector, L.B., & Fortuna, J. (2004). The real stakes of virtual publishing: The transformation of e-biomed into PubMed Central. *Journal of the American Society for Information Science & Technology*, *55*(2), 127-148.

Kurtz, M.J., Eichhorn, G., Accomazzi, A., Grant, C., Demleitner, M., Henneken, E., & Murray, S.S. (2005). The effect of use and access on citations. *Information Processing & Management*, *41*, 1395-1402.

Manuel, K. (2001). The place of e-prints in the publication patterns of physical scientists. *Science & Technology Libraries*, *20*(1), 59-85.

Mikhailov, A.I., Chernyi, A.I., & Giliarevskii, R. (1984). *Scientific communication and informatics*. Arlington: Information Resources.
Mine, S. (2009). The roles and place of arXiv in scholarly communication. *Library and Information Science*, *61*, 25-58.
Moed, H.F. (2007). The effect of 'Open Access' on citation impact: An analysis of arXiv's condensed matter section. *Journal of the American Society for Information Science & Technology*, *58*(13), 2047-2054.
Morris, S. (2003). Open publishing. *Learned Publishing*, *16*, 171-176.
Pinfield, S. (2001). How do physicists use an e-print archive? *D-Lib Magazine*, *7*(12).
Rodriguez, M.A., Bollen, J., & Van de Sompel, H. (2006). The convergence of digital libraries and the peer-review process. *Journal of Information Science*, *32*(2), 149-159.
Swan, A., & Brown, S. (2003). Authors and electronic publishing: What authors want from the new technology. *Learned Publishing*, *16*, 28-33.
Wilson, J.H. (1970). International high-energy physics preprint network emphasizes institutional exchange. *Journal of the American Society for Information Science*, *21*(1), 95-97.

ON THE DETERMINANTS OF RESEARCH PERFORMANCE: EVIDENCE FROM ECONOMIC DEPARTMENTS OF FOUR EUROPEAN COUNTRIES (RIP)

Stelios Katranidis,[1] Theodore Panagiotidis[2] and Costas Zontanos[3]

[1] *katranid@uom.gr*
University of Macedonia, Dept. of Economics, 156 Egnatia Str., 540 06 Thessaloniki (Greece)

[2] *tpanag@uom.gr*
University of Macedonia, Dept. of Economics, 156 Egnatia Str., 540 06 Thessaloniki (Greece)

[3] *zontanos@uom.gr*
University of Macedonia, Library & Information Center, 156 Egnatia Str., 540 06 Thessaloniki (Greece)

Abstract
This paper investigates the research performance of 404 economists working in 17 Economics Departments from 4 countries. We compare two countries from the North of Europe (Belgium and Denmark) with two countries from the South (Greece and Portugal). We differentiate the research performance of their faculty depending on the country they did their PhD and the country of their current affiliation. Based on these, we rank their performance. Furthermore, we employ regression analysis to identify the factors that drive the research performance taking into account both the research environment they have faced while PhD students and as faculty. The most productive economists have a PhD from the US and work in the North.

Conference Topic
Science Policy and Research Evaluation: Quantitative and Qualitative Approaches (Topic 3) and Management and Measurement of Bibliometric Data within Scientific Organizations (Topic 9).

Introduction
There has been an increasing interest on ranking researchers, faculty members, departments, universities and scientific journals. This is evident from the increase in publications on the topic. This paper evaluates the research work of - 404 economists – faculty members of 17 European Economics Departments; 8 of them in Belgium, 2 in Denmark, 3 in Portugal and 4 in Greece. The objectives of this study are: first, by adopting more than one productivity indicators, we evaluate the published research work and make comparisons. Second, we

demonstrate that the place of doctoral studies (Ph.D. origin) affects research performance. Third, we examine the relation between the place of affiliation and research productivity.

Review of the relevant literature

A number of papers related to research evaluation and academic ranking have been published recently. The data they employ are based on the number of papers published in international refereed academic journals and their corresponding citations. The former are globally recognized as the main outlet of scientific work in economics. Many of these studies assess quality via the journals' (Combes, & Linnemer, 2003; Kalaitzidakis, Mamuneas, & Stengos, 2003; Kalaitzidakis, Mamuneas, & Stengos, 2011). Attempts of this kind, i.e. evaluation and rankings of European economists and economic institutions based on this method, include Kalaitzidakis, Mamuneas, & Stengos (1999) on Greek university departments, Guimaraes (2002) on Portuguese, Bauwens (2003) on Belgian, Çokgezen (2006) on Turkish faculty and universities and Lubrano, Bauwens, Kirman, & Protopopescu (2006) where faculties and research performance was compared for 18 European countries. On the other hand, more recently published papers, for Ireland (Ruan, & Tol, 2008; Tol, 2008), for Israel (Ben-David, 2010), for Sweden (Henkerson, & Waldenström, 2011) and for Greece (Katranidis, Panagiotidis, & Zontanos, 2012), rely on bibliometric databases (Web of Science, Google Scholar and Scopus) and consider directly the scientific impact of each paper separately, i.e. according to the times it has been cited.

Besides rankings, some of the above papers proceed even further; for example, to examine the differences in research performance between private and state universities (Çokgezen, 2006), high and low rank academic positions (Ben-David, 2010) and differences in academic advancements due to the country where the doctoral studies have been carried out (Guimaraes, 2001; Katranidis, Panagiotidis, & Zontanos, 2012).

The factors that have appeared in the literature as determinants of research performance include (i) the place and especially the university where the doctoral studies were completed, and (ii) the department where the faculty member serves as a researcher, and both factors in combination with the time that has elapsed since completing the doctoral studies (academic age) (Long, 1978; Long, Bowers, Barnett, & White, 1998).

Our sample includes 404 faculty members from 17 departments. We have determined the sample size using two criteria: (a) at the national level the overall sample size represents 25% of the RePEc registered economists, (b) within that total number, we established a hierarchy of departmental affiliation, whereby we gave priority to the members of higher ranking departments, until reaching the designated ceiling. The departments involved are listed in the appendix.

We calculate the following bibliometric indices: productivity (number of publications per faculty); overall impact (number of citation per faculty) and the rational h^*-index as it has been initially proposed by Hirsch (2005) and modified

by Tol (2008), all of them divided by the research age, i.e. the number of years after the completion of their PhD. These indices are lightening several aspects of what the literature refers to as research performance.
The analysis for each faculty member of all the seventeen economics departments under consideration is based on data retrieved form Scopus (March-May 2012). The faculty members have been identified according to the Website of each department (only economists have been considered). Data on the research work of each faculty member (number of papers, number of citations, and *h**-index) were collected from the Scopus citation database.

Descriptive statistics

Table 1. Research age and faculty distribution according to the PhD origins in %

Country	*Research Age*	*PhD Origins in %*					*Faculty members*
		Dpt. or Inst.	Home	Europe	UK	Overseas	
Belgium	18.01	50.39	19.38	13.18	2.33	14.73	129
Denmark	16.05	59.32	6.78	18.64	1.69	13.56	59
Greece	20.22	9.75	8.54	10.98	39.02	31.71	82
Portugal	15.79	45.92	3.06	19.39	10.20	21.43	98

Table 2. Bibliometric indices, PhD origins by country of affiliation

Country	*Bibliomet. Indices*	*PhD Origins*				*Total*
		Home	Europe	UK	Overseas	
Belgium	p/f	0.61	0.95	-	0.70	0.66
	c/f	3.64	5.69	-	7.82	4.38
	h*	0.23	0.38	-	0.30	0.26
Denmark	p/f	0.66	0.66	-	0.64	0.67
	c/f	4.59	5.38	-	6.17	4.97
	h*	0.27	0.29	-	0.32	0.29
Greece	p/f	0.26	0.25	0.45	0.58	0.41
	c/f	0.58	0.96	1.67	3.44	1.75
	h*	0.12	0.12	0.15	0.24	0.17
Portugal	p/f	0.23	0.25	0.29	0.36	0.27
	c/f	0.33	0.93	1.32	1.84	0.72
	h*	0.09	0.13	0.17	0.17	0.12

Note: p/f: papers per faculty per year, c/f: citations per faculty per year, h is a rational h-index divided by research age.*

Table 1 presents the average research age, the percentage distribution of faculty members according to the country they completed their PhD and the total number of faculty members per country. It is worth mentioning the very low percentage of Greek faculty members having completed their PhD studies in their own country. Moreover, a large part of Greek faculty, almost like their Portuguese colleagues, has completed their studies abroad, mainly in the UK and overseas (mostly in the

US and Canada). These rates, regarding overseas studies for the Belgian and Danish faculty members are much lower and almost negligible for PhD studies in the UK.
Table 2 presents the research performance depending on the place of completing the doctorate studies (PhD origin) and the country of employment (affiliation).

Regression Results

At this section we take the further step to identify the factors that drive the research performance of an economist. There are two important factors that require further examination. On the one hand we have the place that his/her PhD studies took place as a proxy of the training and the research culture that the researcher did face (factor one: research training). On the other, we want to examine the impact of the working environment as the latter will influence the research performance based on the research culture of the department, the interaction with his/her colleagues, the peer pressure and research seminars among other factors (factor 2: work environment). The adopted specification allows us to make at least two comparisons: (i) compare researchers according to whether they did their PhD studies in the US or the country they currently work and (ii) compare researchers that did their PhD in the US and currently are employed in the four different countries under consideration. As a result, we employ the following econometric specification:

$$Research\ Performance = \alpha_1 PhD^{US} * Belgium + \alpha_2 PhD^{US} * Denmark + \alpha_3 PhD^{US} * Greece + \alpha_4 PhD^{US} * Portugal + \beta_1 PhD^{internal} * Belgium + \beta_2 PhD^{internal} * Denmark + \beta_3 PhD^{internal} * Greece + \beta_4 PhD^{internal} * Portugal$$

where Research Performance can be (i) Papers per faculty per year (*p/f*) or (ii) citations per faculty per year (*c/f*) and PhD^{US} is a dummy variable that takes the value of 1 if the research got his/her PhD from the US. *Belgium*, *Denmark*, *Greece* and *Portugal* are dummy variables for the country that the researcher is currently employed and $PhD^{internal}$ if the researcher got his/her PhD from the same country that he/she is employed.
The purpose of this study is twofold; compare α_1 to α_2, α_3 or α_4 etc and compare α_1, with β_1, α_2 with β_2 etc. If $\alpha_1 + \alpha_2 > \alpha_3 + \alpha_4$, the working environment in the north (Belgium and Denmark) is superior to the one in the south (Greece and Portugal) given the researcher has a similar training.
Table 3 presents the regression outcome for the two alternative specifications (one with *p/f* as dependent variable and one with *c/f*). In the former, we observe that all coefficients are statistically significant at the 5% level. The ranking that emerges for the US PhD holders is: Belgium, Greece, Denmark and Portugal and for the internal PhD holders: Denmark, Belgium, Portugal and Greece. For the papers per year per faculty, it emerges that for someone with a PhD from the US, the average performance per year is 0.957 if currently working in Belgium, 0.889 in Greece, 0.618 in Denmark and 0.452 in Portugal. For the citations per year:

11.9 if one holds a US PhD and works in Belgium, 5.12 in Greece, 4.18 in Denmark and 3.5 in Portugal. With regard to the comparison between north vs south: = 0.957+0.618 = 1.575 > 1.341 = 0.889+0.452= $\hat{\alpha}_3 + \hat{\alpha}_4$. The latter implies that the working environment in the north stimulates research more compared to the south (although this difference is not statistically significant *p*-value 0.6). These results are not qualitatively different when we consider alternative measures of research performance (results available upon request).

Table 3. OLS Coefficients

Dependent Variable	*Papers per Year*	*Citations per Year*
PhD_US*BE_DUMMY	0.957 ***	11.921 **
	3,229	2,169
PhD_US*DK_DUMMY	0.618 **	4.184 **
	2,402	2,075
PhD_US*GR_DUMMY	0.889 ***	5.123 ***
	4,093	3,578
PhD_US*PT_DUMMY	0.452 ***	3.501 **
	5,346	2,505
PhD_BE*BE_DUMMY	0.814 ***	6.025 ***
	9,648	7,540
PhD_DK*DK_DUMMY	0.946 ***	6.928 ***
	7,205	6,457
PhD_GR*GR_DUMMY	0.454 **	1.280 **
	2,463	2,235
PhD_PT*PT_DUMMY	0.513 ***	2.537 **
	2,949	2,249

*Note: t-statistics below the estimate coefficients, ***, ** and * significance at the 1%, 5% and 10 % respectively.*

The analysis so far focused on the average researcher (mean response). However, it might be more useful to look at the median researcher as this is more representative. For this, we employ quantile regression (QR) that investigates the median response rather than the average response (see for instance Koenker, & Bassett, 1978). The results for the same specification are presented in Table 4.
The research productivity as it is measured by the alternative specifications is summarized in the following table where each coefficient is divided by the max coefficient. As it is evident from the relative numbers in Table 5 OLS overestimate the research productivity of those who are based in Greece and Portugal relative to the best (the best is researchers based in Belgium for US PhD holders and Denmark for local PhD holders). The QR coefficients are lower compared to the OLS ones. The latter can be explained by the fact that the median researcher is more representative than the average one given the outliers that exist in these countries. The most notable difference emerges when one compares the productivity of the median researcher from Denmark with PhD from the same

country and the respective median researcher from Portugal (or Greece). We observe that the latter has only 4% (or 6% in the case of Greece) of the citations of the median researcher from Denmark.

Table 4. Quantile regression coefficients (median)

Dependent Variable	*Papers per Year*	*Citations per Year*
PhD_US*BE_DUMMY	0.833 ***	4.629 *
	4,501	1,678
PhD_US*DK_DUMMY	0,250	0,500
	1,498	2,075
PhD_US*GR_DUMMY	0.474 ***	1.857 **
	4,399	2,413
PhD_US*PT_DUMMY	0.286 ***	0,813
	2,643	1,354
PhD_BE*BE_DUMMY	0.643 ***	2.250 ***
	4,943	4,108
PhD_DK*DK_DUMMY	0.750 ***	3.500 ***
	6,049	4,149
PhD_GR*GR_DUMMY	0.222 **	0,222
	2,525	0,388
PhD_PT*PT_DUMMY	0.214 ***	0,154
	3,805	0,490

Table 5. Relative productivity

Relative Productivity p/f			*Relative Productivity c/f*		
US PhD	*OLS*	*QR*	*US PhD*	*OLS*	*QR*
BE	1,00	1,00	BE	1,00	1,00
DK	0,65	0,30	DK	0,35	0,11
GR	0,93	0,57	GR	0,43	0,40
PT	0,47	0,34	PT	0,29	0,18
Local PhD			*Local PhD*		
DK	1,00	1,00	DK	1,00	1,00
BE	0,86	0,86	BE	0,87	0,64
GR	0,48	0,30	GR	0,19	0,06
PT	0,54	0,29	PT	0,37	0,04

Conclusions

This paper examines the research performance of 404 economists working in 17 Economics Department from 4 European countries. Comparisons are made between two countries from the North of Europe (Belgium and Denmark) with two countries from the South (Greece and Portugal).The analysis for the research performance of their faculty takes into account the country they did their PhD on the one hand and their current affiliation on the other. We provide a ranking of

their absolute and relative performance based on these two criteria. Furthermore, we identify the factors that drive the research performance by employing regression analysis and quantile regression analysis based on both PhD origins and current country affiliation. In terms of productivity (impact) the following order emerges: 1 (1) PhD from the US and works in Belgium, 2 (2) PhD from Denmark and works in Denmark, 3 (4) PhD from the US and works in Greece, 4 (3) PhD from Belgium and works in Belgium, 5 (5) PhD from the US and works in Denmark, 6 (7) PhD from Portugal and works in Portugal, 7 (8) PhD from Greece and works in Greece and 8 (6) PhD from the US and works in Portugal.

References

Bauwens, L. (2003, May 31). Economic research in Belgian universities. Retrieved from: http://www.core.ucl.ac.be/econometrics/Bauwens/rankings/rankings.htm

Ben-David, D. (2010). Ranking Israel's economists, *Scientometrics, 82,* 351-364.

Çokgezen, M. (2006). Publication performance of economists and economics departments in Turkey (1999-2003). *Bulletin of Economic Research, 58,* 253-265.

Combes, P.P., & Linnemer, L. (2003). Where are the economists who publish? Publication concentration and rankings in Europe based on cumulative publications. *Journal of the European Economic Association, 1,* 1250-1308.

Guimarães, P. (2002). The state of Portuguese research in economics: An analysis based on publications in international journals. *Portuguese Economic Journal, 1,* 3-25.

Henkerson, M., & Waldenström, D. (2011). How should research performance be measured? A study of Swedish economists. *Manchester School, 79,* 1139-1156.

Hirsch, J.E. (2005). An index to quantify an individual's scientific research output. *Proceedings of the National Academy of Sciences of the United States of America, 102,* 16569-16572.

Kalaitzidakis, P., Mamuneas, T.P., & Stengos, T. (1999). Ranking of economics departments among Greek-speaking institutions. *Economia, 3,* 70-75.

Kalaitzidakis, P., Mamuneas, T.P., & Stengos, T. (2003). Rankings in academic journals and institutions in economics. *Journal of the European Economic Association, 1,* 1346-1366.

Kalaitzidakis, P., Mamuneas, T.P., & Stengos, T. (2011). An updated ranking of academic journals in economics. *Canadian Journal of Economics, 44,* 1525-1538.

Katranidis, S., Panagiotidis, T., & Zontanos, C. (2012). An evaluation of the Greek universities' economics departments. *Bulletin of Economic Research.* Advance online publication. doi: 10.1111/j.1467-8586.2012.00434.x

Koenker, R., & Bassett, G. (1978). Regression quantiles. *Econometrica, 46,* 33-50.

Long, J.S. (1978). Productivity and academic position in the scientific career. *American Sociological Review, 43,* 889-908.
Long, R.G., Bowers, W.P., Barnett, T., & White, M.C. (1998). Research productivity of graduates in management: Effects of academic origin and academic affiliation. *Academy of Management Journal, 41,* 704-714.
Lubrano, M., Bauwens, L., Kirman, A., & Protopopescu, C. (2003). Ranking economics departments in Europe: A statistical approach. *Journal of the European Economic Association, 1,* 1367-1401.
Ruan, F., & Tol, R.S.J. (2008). Rational (successive) h-indices: An application to economics in the Republic of Ireland. *Scientometrics, 75,* 395-405.
Tol, R.S.J. (2008). A rational, successive g-index applied to economics departments in Ireland. *Journal of Informetrics, 2,* 149-155.

Appendix

Belgium: 1) École des Sciences Économiques de Louvain. Université Catholique de Louvain, 2) Department of Economics. Faculteit Economie en Bedrijfswetenschappen. Katholieke Universiteit Leuven, 3) European Centre for Advanced Research in Economics and Statistics (ECARES) & Département d'Économie Appliquée (DULBEA). Solvay Brussels School of Economics and Management. Université Libre de Bruxelles, 4) Economics Department. HEC École de Gestion. Université de Liège, 5) Department of Economic Science. Faculté des Sciences Économiques, Sociales et de Gestion (FSESG). Facultés Universitaires Notre-Dame de la Paix (Namur), 6) Dpts. General, Financial & Social Economics. Faculteit Economie en Bedrijfskunde. Universiteit Gent, 7) Dpt. Of Economics. Faculteit Toegepaste Economische Wetenschappen. Universiteit Antwerpen, 8) Centre de Recherche en Économie (CEREC). Facultés Universitaires Saint-Louis
Denmark: 1) Institut for Økonomi, Aarhus Universitet, 2) Københavns Universitet. Økonomisk Institut
Greece: 1) Dpt. of Economics, Athens University of Economics and Business, 2) Dpt. of Economics, University of Crete, 3) Dpt. of Economics, University of Macedonia, 4) Dpt. of Economics, University of Piraeus
Portugal: 1) Universidade Nova de Lisboa. Faculdade de Economia, 2) Instituto Superior de Economia e Gestão (ISEG). Dpt. of Economics. Universidade Técnica de Lisboa, 3) Grupo de Economia. Faculdade de Economia. Universidade do Porto

OPEN DATA AND OPEN CODE FOR BIG SCIENCE OF SCIENCE STUDIES

Robert P. Light, David E. Polley, and Katy Börner

lightr@indiana.edu, dapolley@indiana.edu, katy@indiana.edu
Cyberinfrastructure for Network Science Center, School of Library and Information Science, Indiana University, Bloomington, IN, USA

Abstract

Historically, science of science studies were/are performed by single investigators or small teams. As the size and complexity of data sets and analyses scales up, a "Big Science" approach (Price, 1963) is required that exploits the expertise and resources of interdisciplinary teams spanning academic, government, and industry boundaries. Big science of science studies utilize "big data", i.e., large, complex, diverse, longitudinal, and/or distributed datasets that might be owned by different stakeholders. They apply a systems science approach to uncover hidden patterns, bursts of activity, correlations, and laws. They make available open data and open code in support of replication of results, iterative refinement of approaches and tools, and education. This paper introduces a database-tool infrastructure that was designed to support big science of science studies. The open access Scholarly Database (SDB) (http://sdb.cns.iu.edu) provides easy access to 26 million paper, patent, grant, and clinical trial records. The open source Science of Science (Sci2) tool (http://sci2.cns.iu.edu) supports temporal, geospatial, topical, and network studies. The scalability of the infrastructure is examined. Results show that temporal analyses scale linearly with the number of records and file size, while the geospatial algorithm showed quadratic growth. The number of edges rather than nodes determined performance for network based algorithms.

Conference Topic

Old and New Data Sources for Scientometric Studies: Coverage, Accuracy and Reliability (Topic 2), Visualisation and Science Mapping: Tools, Methods and Applications (Topic 8) and Open Access and Scientometrics (Topic 10)

Introduction & Related Work

Many science of science studies use heterogeneous datasets and advanced data mining and visualization algorithms advance our understanding of the structure and dynamics of science. The quality of results depends on the quality and coverage of the data used. Data cleaning and preprocessing can easily consume 80 percent or more of the overall project effort and budget. As the number of data records grows, different types of tools and expertise are required to handle the data. MS Excel can load a maximum of 1,048,576 rows of data by 16,384 columns per sheet. MS Access file sizes cap at 2 gigabytes, including indices, forms, and macros along with the data. Larger datasets need to be stored in a database designed with scalability in mind. As the diversity of datasets increases,

the structures of different datasets need to be aligned. As data covers more and more years, dealing with format changes becomes necessary. Many studies require extensive preprocessing and augmentation of the data, such as identification of unique records or record values, geocoding of records in preparation for geospatial analysis, or the extraction of networks for network studies. For many researchers, the effort to compile ready-to-analyze-and-visualize data is extremely time consuming and challenging and sometimes simply insurmountable.

Many datasets relevant for science of science studies, e.g., papers, patents, grants, clinical trials, are freely available by different providers. However, they are stored in separate silos with diverse interfaces of varying usability that deliver data in many different formats. Research projects seeking to use one or many of these data sources face major data access, integration, and unification challenges. Indiana University's Scholarly Database (SDB), originally launched in 2005, makes over 26 million scholarly records freely available via a unified interface and in data formats that are easy to use and well documented. In the last four years, SDB has answered thousands of queries and delivered millions of records to users around the globe. The 2012 update to the SDB improves the quality of data offered and integrates new humanities and clinical trial datasets.

Equipped with high quality, high coverage data in standard data formats, tools that scale in terms of the number of records that can be read and processed are needed to truly make sense of big data (Robertson, Ebert, Eick et al., 2009). While most tools work well for micro and meso level studies (up to 100,000 records), few scale to macro level big-data studies with millions or even billions of records. Another type of scalability relates to the ease of usage and ease of interpretation of big data visualizations. How to best communicate temporal trends or burst of activity over a 100 year time span? How to depict the geospatial location of millions of records in a scalable fashion? Can the topical evolution of massive document datasets be communicated to a general audience? Most visualizations of million node networks resemble illegible spaghetti balls—do advanced network analysis algorithms scale and help to derive insights?

Frequently, different types of analysis have to be applied to truly understand a natural, social, or technological system. Examples are temporal studies that answer WHEN questions, geospatial studies that answer WHERE questions and draw heavily on research in cartography, topical studies that use linguistic analysis to answer WHAT questions, and network studies that employ algorithms and techniques developed in social sciences, physics, information science and other domains to answer WITH WHOM questions. However, most existing systems support only one general type of analysis and visualization and many require programming skills. For example, four of the top 20 data visualization tools listed by *.net* in September of 2012 support charts and graphs while six

support geospatial maps exclusively (Suda, 2012). Only the D3 (Data-Driven Documents) and Raphaël JavaScript libraries, the Google Chart API, and R support a larger array of charts, graphs, and maps yet all three require programming or scripting skills that most users do not possess. Excel might be the only tool on the list that can be used by a large number of non-programmers. A listing of tools commonly used in science of science studies can be found at http://sci2.wiki.cns.iu.edu/display/SCI2TUTORIAL/8.2+Network+Analysis+and+Other+Tools but most support a very limited range of workflows (Cobo, López-Herrera, Herrera-Viedma et al., 2011).

This paper presents a database-tool infrastructure that applies a divide-and-conquer approach to support big science of science studies. It combines an online database supporting bulk download of data in easy to process formats with a plug-and-play tool to read, clean, interlink, mine, and visualize data using easy to manipulate graphical user interfaces.

The remaining paper is organized as follows: The next two sections present the database and tool functionalities. Subsequently, we test and discuss their scalability. We conclude the paper with a discussion of the presented work and an outlook to future work.

The Scholarly Database (SDB)

The Scholarly Database was created in 2005 to provide researchers and practitioners easy access to various datasets offered by different publishers and agencies (LaRowe, Ambre, Burgoon et al., 2009). The Scholarly Database is implemented using PostgreSQL 8.4, a free and open source relational database management system. Since the introduction of version 8.1, PostgreSQL developers have been focused on improving the scalable performance of the system and this software is now employed by many companies to provide large-scale data solutions, including Yahoo!, Sony Online and Skype. Today, the Scholarly Database provides easy access to paper, patent, grant, and clinical trials records authored by 13.8 million people in 208 countries (some, such as Yugoslavia, no longer in existence), interlinked by 58 million patent citation links, and over 2.5 million links connecting grant awards to publications and patents. As of November 2012, the SDB features over 26 million records from MEDLINE (19,039,860 records spanning from 1865-2010), USPTO patents (4,178,196, 1976-2010), NIH awards (2,490,837, 1972-2012), NSF awards (453,687, 1952-2010), NEH awards (47,197, 1970-2012) Clinical Trials (119,144, 1900-2012).

Unique features of SDB comprise:

- *Open Access:* The SDB is composed entirely of open data so there are no copyright or proprietary issues for the researcher to contend with in its use. Data is provided to researchers free of charge.

- *Ease of Use:* Simple user interfaces provide a one-stop data access experience making it possible for researchers to focus on answering their questions, rather than spending much time on parsing, searching, and formatting data.
- *Federated Search:* By aggregating the data into a single environment, SDB offers a federated search environment powered by a Solr core. Users can search one, some, or all of the available datasets over some or all years using the same set of terms and get a combined set of results that are ranked by relevance.
- *Bulk Download:* Most databases do not support downloads and those that do only permit access to a limited number of records. SDB supports bulk download of data records; data linkages—co-author, patent citations, grant-paper, grant-patent; burst analysis files. Users are granted a base number of downloads by default to prevent abuse of the system, but this number can be extended by request without charge.
- *Unified File Formats:* SDB source data comes in different file formats. NIH funding data is stored in flat files; clinical trials are offered in XML, while patents come in a variety of formats, depending on the year. Old patents come in a fixed width data format while newer patents are provided in XML. Much time and effort was spent to normalize this data into easy-to-use file formats, e.g., comma-delimited tables for use in spreadsheet programs and common graph formats for network analysis and visualization.
- *Well-Documented:* SDB publishes data dictionaries for every dataset offered. Information on data provenance, table structure, data types, and individual field comments are available. In addition, the SDB offers a set of small sample files, giving researchers an easily usable test-bed for working out their algorithms before committing to analysis of a larger set.

The SDB Wiki (http://sdb.wiki.cns.iu.edu) provides more information including a user guide, information on each dataset, and release notes.

The Science of Science (Sci2) Tool

The Science of Science (Sci2) tool is a modular toolset specifically designed for the study of science. It supports the temporal, geospatial, topical, and network analysis and visualization of scholarly datasets at the micro (individual), meso (local), and macro (global) levels, see screenshot in Figure 1, general workflow in Figure 2 and specific workflows discussed in the scalability tests section.

The tool's OSGi/CIShell core architecture makes it possible for domain scientists to contribute new algorithms written in a variety of programming languages using a plug-and-play macroscope approach (Börner, 2011).

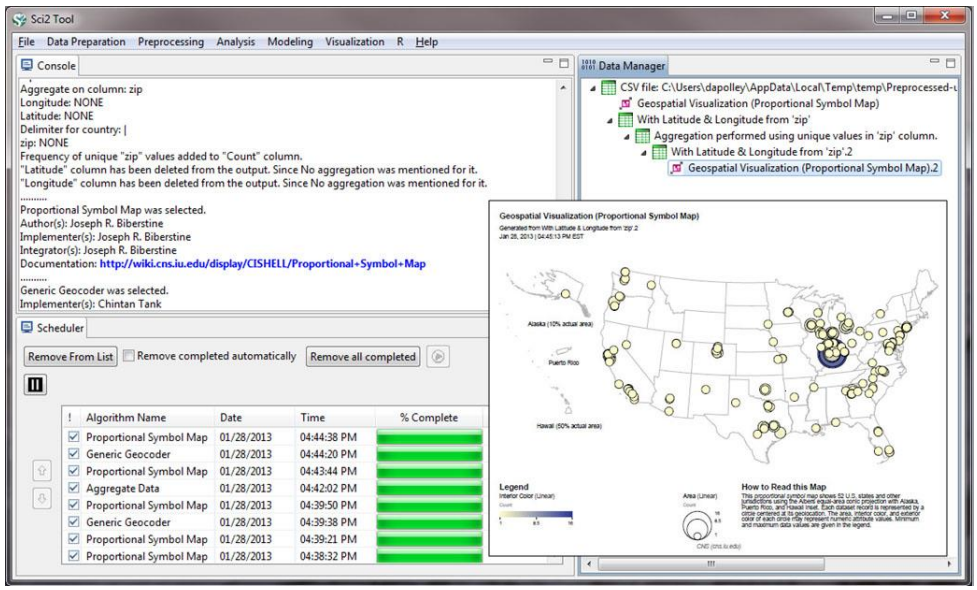


Figure 1: Sci2 tool user interface with proportional symbol map visualization.

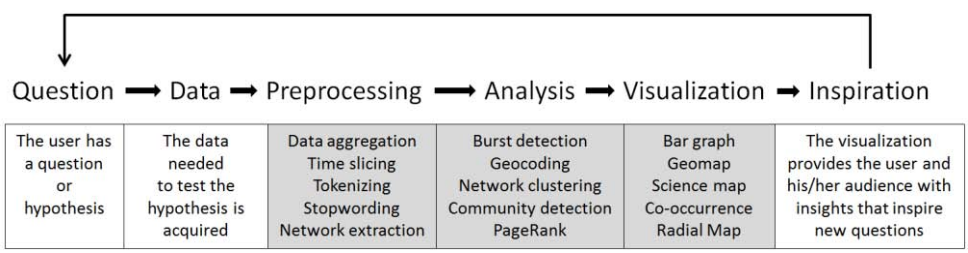


Figure 2: General Sci2-based visualization creation workflow (tool-specific tasks in gray).

Category	Algorithms	Examples
Acquisition	5	Google Citation User ID Search Algorithm
Data Preparation	13	Extract Co-Occurrence Network
Preprocessing	22	Slice Table by Time, Extract ZIP Code
Analysis	47	K-Nearest Neighbor, Burst Detection
Modeling	4	Watts-Strogatz Small World, TARL
R	4	Create an R Instance, Send a Table to R
Visualization	17	Choropleth Map, Bipartite Network Graph
Total	**112**	

Figure 3: Sci2 algorithm summary tables.

As of November 2012, the Sci2 tool has 171 algorithms, 112 of which are visible to the user (see Figure 3) written in Java, C, C++, and Fortran. In addition, a number of tools (Gnuplot, Guess, and Cytoscape) were implemented as plugins and bridges to R and to Gephi were created, allowing the seamless use of different tools. The Sci2 user interface and sample map is shown in Figure 1.

Unique features of Sci2 comprise:

- *Open Source*: Anybody can examine the source code and advance it.
- *Extensive use of well-defined reference systems*: To improve readability and to support interpretation, Sci2 uses a number of carefully designed reference systems, see Figure 4. Each comes with a title, legend, and a brief "How to read this visualization" section that provides further details, e.g., on used geospatial projections.
- *Interactivity*: While visualizations of small datasets can be explored interactively, visualizations of big data are rendered into Postscript files that can be converted to pdf files and examined using pan and zoom as well as filtered, e.g., by searching for specific text in the display.
- *Workflows*: All user actions are recorded in a log file to ensure proper documentation and easy replicability of workflows that might comprise 15-20 analysis and visualization algorithms with a range of parameter settings.
- *Online documentation*: All Sci2 plugins as well as major workflows are documented in the Sci2 Wiki (http://sci2.wiki.cns.iu.edu) together with release notes.

Scalability Tests

To demonstrate the scalability of the database and tool, tests were performed using synthetic datasets with pre-defined properties generated in Python and datasets retrieved from the Scholarly Database. All four types of analysis supported by Sci2 were tested: temporal analysis, geospatial analysis, topical analysis, and network analysis. Initially, we identified workflows indicative of these four main types of analysis. From there, we broke down each workflow into the specific steps (algorithms) involved in the workflow, starting with loading the data and ending in visualization. For each algorithm, e.g., data reader, analysis, visualization, we measured (in seconds) the length of time it took for an algorithm to finish processing. We considered the start of the algorithm to be the point at which the user inputs his or her parameters (where applicable) and then executes the algorithm. We considered all algorithms to be finished when the associated data files appeared in the Data Manager and were displayed as complete in the Scheduler. For each test, we calculated the average for 10 trials. Between trials, we closed down Sci2 in order to minimize any adverse effects of residual memory. Tests were performed on a common system: an Intel(R) Core(TM) Duo CPU E8400 3.00GHz processor and 4.0GB of memory running a 64bit version of

Windows 7 and a 32bit version of Java 7. Memory allotted to Sci2 was extended to 1500 MB.

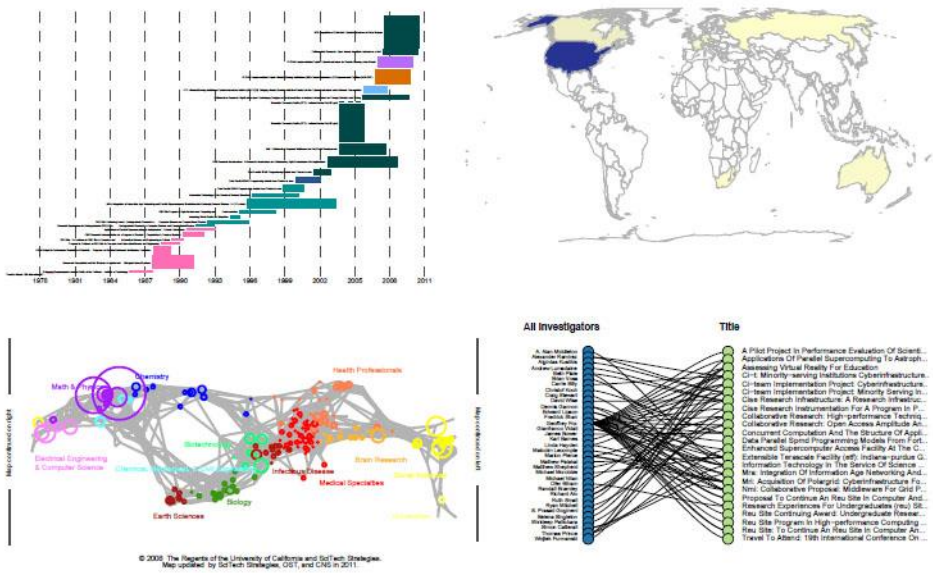


Figure 4: Exemplary reference systems supported by Sci2 including Temporal Bar Graph (top, left), Choropleth map (top, right), UCSD science map (bottom, left), bimodal network visualization (bottom, right) Full versions available at http://wiki.cns.iu.edu/display/SCI2TUTORIAL/1+Introduction

Records	Columns	Size (MB)	Load Time (sec)	SD (sec)
50,000	2	0.48	0.72	0.06
100,000	2	0.95	1.08	0.04
500,000	2	4.77	3.75	0.05
1,000,000	2	9.54	7.14	0.14
1,500,000	2	14.31	10.26	0.08
2,000,000	2	19.07	13.26	0.17
2,500,000	2	23.84	16.47	0.13
50,000	25	5.96	3.56	0.07
100,000	25	11.92	6.44	0.05
500,000	25	59.61	29.62	0.91
1,000,000	25	119.21	122.36	0.64
1,500,000	25	178.81	-TF*	
2,000,000	25	238.42	-TF*	
2,500,000	25	298.02	-TF*	

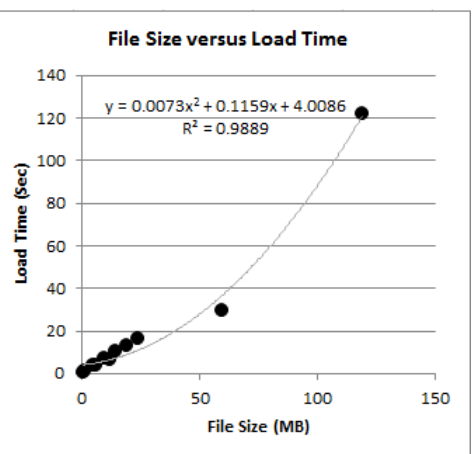


Figure 5: Comparison of load times, measured in seconds, across standardized datasets, tabulated (left) and plotted with quadratic regression line (right).

File Loading

Synthetic data was used to measure how file loading times vary in terms of number of records and length of individual record in bytes. Two series of datasets were generated, one with only two rows, a small integer, and a short string and one with 25 rows, a small integer and 24 short strings, each with increasing numbers of rows. Average loading times over ten trials are given in Figure 5. The three largest datasets did not load but returned a Java heap space error (-TF*). At first glance, there seems to exist a direct relationship between file size and loading time (R^2 = 0.9384), a closer look at the plot of size versus time reveals that a quadratic regression line has a noticeably better fit (R^2=0.9889). This is likely a result of the tool having to devote resources to file management that would otherwise be available for completing functions more efficiently.

Next, SDB data prepared for usage in science of science workflows was read comprising

- NIH data at 3.4GB, NSF data at 489MB, NIH data at 139MB, and NEH data at 12.1MB data prepared for temporal analysis.
- Data from NIH, NSF, MEDLINE, UPSTO, and Clinical Trials at 11.5 MB and MEDLINE data at 1GB to be used in geospatial analysis.
- MEDLINE data at 514KB for topical analysis.
- NSF data at 11.9MB and UPSTO data at 1.04GB network analysis.

Average load times measured across ten trials are shown in Table 1. The three largest datasets, would not load but returned a Java heap space error (-TF*).

Table 1: Comparison of load times, measured in seconds, across nine different datasets.

Dataset	Size	Number of Records	Mean	Standard Deviation	Minimum	Maximum
NIH (year, title, abstract)	3.4GB	2,490,837	-TF*			
USPTO (patent, citations)	1.04GB	57,902,504	-TF*			
MEDLINE (geospatial)	1.0GB	9,646,117	-TF*			
NSF (year, title, abstract)	489MB	453,740	64.54	0.991	63.2	65.9
NIH (title, year)	139MB	2,490,837	83.86	1.32	82.3	85.6
NEH (year, title, abstract)	12.1MB	47,197	2.05	0.070	1.9	2.1
NSF (co-author network)	11.9MB	341,110	4.52	0.063	4.4	4.6
Combined geo-spatial	11.5MB	11,549	1.91	0.056	1.8	2.0
MEDLINE journals	0.5MB	20,775	0.44	0.096	0.3	0.6

Temporal Studies ("When")

To test the scalability of temporal analysis within Sci2 we selected the *Burst Detection* algorithm as described by Kleinberg (2003). To test this in a standardized fashion, we generated a randomized set of years from 1980 to 2000, assigning each year a distribution of short strings to test the accuracy of the

algorithm. We then calculated the average time, minimum time, and the maximum time it took the *Burst Detection* algorithm to complete across ten trials. In all cases, the algorithm was able to detect a pre-programmed burst of a word over a short time frame.

A look at the table and graph in Figure 6 shows linear growth with number of records that holds equally true with file size. It is possible that with larger files, this may begin to show the same quadratic tendency as the file loading, but 2.5 million records was the largest file loaded. The data does illustrate that, barring resource exhaustion issues, Sci2 runs this algorithm in a linear timescale.

Records	Size (MB)	Run Time (sec)	SD (sec)
50,000	0.48	0.75	0.07
100,000	0.95	1.03	0.05
500,000	4.77	3.55	0.07
1,000,000	9.54	6.67	0.07
1,500,000	14.31	9.76	0.18
2,000,000	19.07	13.15	0.17
2,500,000	23.84	15.73	0.22

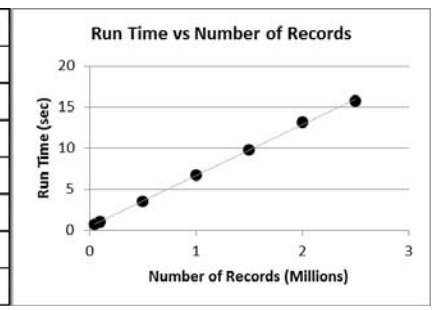


Figure 6: Comparison of Burst Detection run times, measured in seconds, across standardized datasets, tabulated (left) and plotted (right).

We then conducted a burst analysis of the title fields for NIH, NSF, and NEH grant data. The NSF and NEH datasets contain three columns: title, abstract, and year. The NIH data contains only two columns: title and year. The NIH grant data set is the largest at 139MB and 2,490,837 records, followed by the NSF grant data at 489MB and 453,740 records, and finally the NEH grant data at 12.1MB with 47,197 records. In order to obtain accurate results with the *Burst Detection* algorithm we had to normalize the title text with the *Lowercase, Tokenize, Stem, and Stopword Text* algorithm prior to running the *Burst Detection* algorithm, a step not necessary with the synthetic data since it was optimized for burst analysis. Due to the number of records in the NIH dataset, the *Lowercase, Tokenize, Stem, and Stopword Text* algorithm failed to terminate and as a result the *Burst Detection* algorithm was not tested with this dataset (-NT*).

Table 2: Temporal Analysis Algorithm Run Time in seconds.

Burst Detection						
Dataset	**Size**	**Rows**	**Mean**	**SD**	**Min**	**Max**
NSF	489 MB	453,740	13.64	0.648	12.9	14.8
NIH	139 MB	2,490,837	-NT*			
NEH	12.1 MB	47,197	1.57	0.094	1.4	1.7

Geospatial Studies ("Where")

In order to test Sci2 performance for geomapping, randomized datasets with lists of U.S. cities and associated longitude and latitude, were generated. There was only one distinct step (algorithm) involved in this geospatial workflow: visualizing the geolocated data with the *Proportional Symbol Map* (Biberstine, 2012), see U.S. geomap in Figure 2. We projected this on a map of the United States, as this data set only included locations within the U.S. Average run times are shown in Figure 7. Like with file loading, the Proportional Symbol Map data is better fit by a quadratic model (R^2 of 0.997 as opposed to 0.9834 for a linear fit).

Records	Size (MB)	Run Time (sec)	SD (sec)
50,000	1.82	6.26	0.25
100,000	3.66	8.86	0.45
500,000	18.71	22.71	2.00
1,000,000	37.52	44.37	5.21
1,500,000	56.81	70.73	2.15
2,000,000	76.09	92.93	5.63
2,500,000	95.38	134.69	2.78

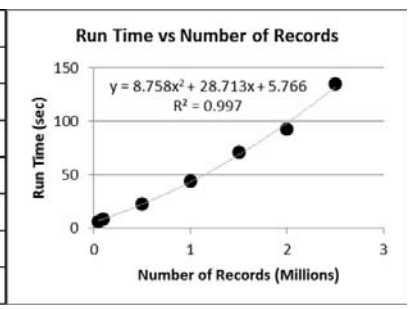


Figure 7: Comparison of Proportional Symbol Map run times, measured in seconds, across standardized datasets.

Next, 11,848 SDB records related to gene therapy funding (NIH, NSF), publications (MEDLINE), patents (USPTO), and clinical trials were loaded and the *Proportional Symbol Map* was used to display the geocoded data. Exactly 299 records had no or incomplete geolocation data and were removed resulting in 11,549 rows at 11.5MB. The run time, at 4.37 sec is lower than predicted by the model (6.11 sec), implying that the quadratic model may not perfectly describe the run time, particularly with smaller sets.

Table 3: Geospatial Analysis Algorithm Run Time in seconds.

Algorithm 1: Proportional Symbol Map						
Dataset	**Size**	**Rows**	**Mean**	**SD**	**Min**	**Max**
Pre-located	11.5 MB	11,549	4.37	0.125	4.2	4.6

Topical Studies ("What")

The Sci2 tool supports the generation of science map overlays. Specifically, it uses the UCSD map of science and classification system (Börner, Klavans, Patek et al., 2012), a visual representation of 554 sub-disciplines within 13 disciplines of science and their relationships to one another, see lower left map in Figure 2. This basemap is then used to show the result of mapping a data set's journals to the underlying subdiscipline(s) those journals represent (Biberstine, 2011).

Mapped subdisciplines are shown with node sizes relative to the number of articles matching journals and color is based on the discipline as defined in the basemap. To create a standardized dataset, random lists of valid journal names were generated. The number of records and run time results are tabulated in plotted in Figure 8. Linear and quadratic models fit about equally well, but both show that the intercept is about 1.5 seconds, more than half of the run time for all but the largest sets. This stands to reason as the lookup tables must be loaded and accessed regardless of the size of the dataset being used.

Records	Size (MB)	Run Time (sec)	SD (sec)
50,000	1.33	1.59	0.13
100,000	2.67	1.58	0.09
500,000	13.79	1.89	0.09
1,000,000	27.66	2.25	0.07
1,500,000	42.02	2.92	0.16
2,000,000	56.40	3.19	0.03
2,500,000	70.77	3.93	0.26

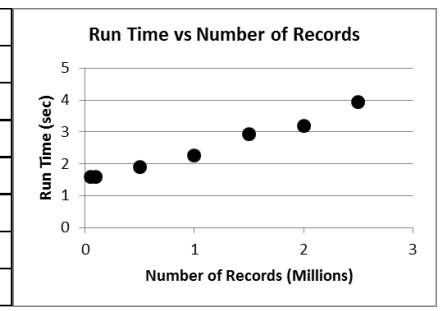


Figure 8: Comparison of UCSD Map of Science Generation run times, measured in seconds, across standardized datasets.

Next, MEDLINE data was obtained from SDB including all 20,773 journals indexed in MEDLINE and the number of articles published in those journals. Average *Map of Science via Journals* run times are given in Table 4.

Table 4: Topical Visualization Algorithm Run Time in seconds.

Algorithm 1: Map of Science via Journals						
Dataset	**Size**	**Rows**	**Mean**	**SD**	**Min**	**Max**
MEDLINE journals	514 KB	20,773	7,84	0.096	7.7	8.0

Network Studies ("With Whom")

Sci2 supports the extraction of diverse network types. The *Extract Directed Network* algorithm (Alencar, 2010) accepts tabular data and constructs a directed network from entities in the specified source column to entities in the specified target column. Run times across ten trials for networks with different numbers of nodes and edges are shown in Figure 9. As to be expected, there is a direct linear relationship between the number of edges and the run time.

Next we retrieved from the SDB all 6,206 USPTO patents that cite patents with numbers 591 and 592 in the patent number field. We ran the *Extract Directed Network* algorithm, creating a network pointing from the patent numbers to the numbers those patents reference in the dataset and results are given in Table 5. While the scalability of Sci2 third-party visualization tools such as GUESS,

Cytoscape, and Gephi do not pertain to Sci2 in a direct way, we were interested to understand their scalability. Neither Cytoscape nor GUESS were capable of rendering the network in a Fruchterman-Reingold layout, while Gephi loaded the network in 2.1 seconds and rendered it in about 40 seconds (the actual process in Gephi is non-terminating, but this was the time to a reasonably defined network). Gephi is able to achieve higher performance due to its ability to leverage GPUs in computing intensive tasks.

Records	% Conn	Edges	Size (MB)	Run (sec)	SD (sec)	Records	% Conn	Edges	Size (MB)	Run (sec)	SD (sec)
500	**2**	**5,000**	0.017	1.13	0.05	**250**	**50**	**31,250**	0.124	1.86	0.05
500	**5**	**12,500**	0.045	1.44	0.07	**500**	**50**	**125,000**	0.546	5.89	0.1
500	**10**	**25,000**	0.093	1.92	0.04	**1,000**	**50**	**500,000**	2.28	20.74	0.12
500	**25**	**62,500**	0.247	3.46	0.08	**1,500**	**50**	**1,125,000**	5.21	45.28	0.44
500	**50**	**125,000**	0.546	5.89	0.1	**2,000**	**50**	**2,000,000**	9.33	79.41	0.62

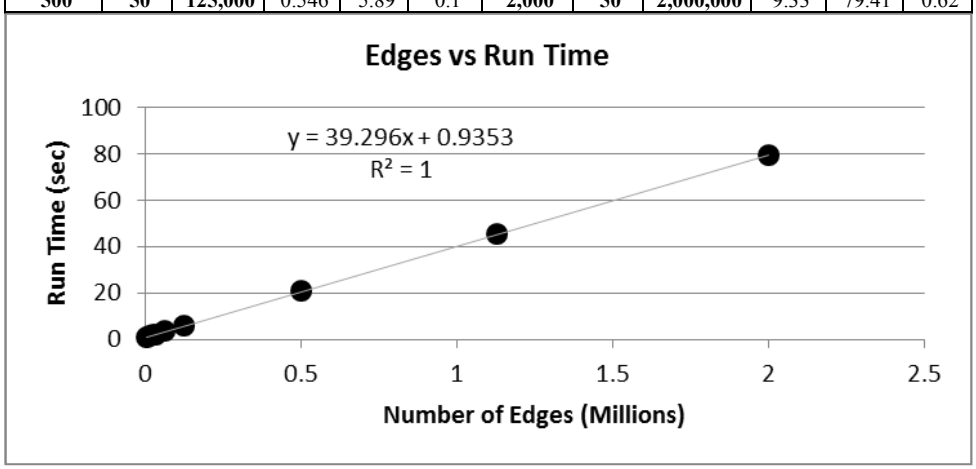


Figure 9: Average Directed Network Extraction run times, measured in seconds versus the number of edges in the dataset, across standardized datasets, tabulated with varying connectivity (left) and number of nodes (right) (top) and plotted (below).

Table 5: Network Analysis Algorithm Run Time in seconds.

Algorithm 1: Extract Co-Occurrence Network							
Dataset	**Size in MB**	**Nodes**	**Edges**	**Mean**	**SD**	**Min**	**Max**
U.S. Patent References	0.147	12,672	7,940	7.88	0.103	7.7	8.1

Discussion and Future Work

This paper introduced and examined the scalability of a database-tool infrastructure for big science of science studies. SDB relational database functionality was exploited to store, retrieve, and preprocess datasets. Subsequently, the data were processed using the Sci2 Tool. The scalability of this approach was tested for exemplary analysis workflows using synthetic

and SDB data. Techniques used were similar to those employed in testing the performance of web-native information visualizations (Johnson & Jankun-Kelly, 2008). Most run-times scale linearly or exponentially with file size. The number of records impacts run-time more than file size. Files larger than 1.5 million records (synthetic data) and 500MB (SDB) cannot be loaded and hence not be analyzed. Run times for rather large datasets are commonly less than 10 seconds. Only large datasets combined with complex analysis require more than one minute to execute.

A forthcoming paper will compare the runtime of Sci2 with other tools that have similar functionality, e.g., TEXTrend or VOSViewer for topical analysis and visualization; CiteSpace, Leydesdorff's Software, DynaNets, SISOB, Cytoscape, and Gephi for network analysis and visualization, see below and (Cobo, López-Herrera, Herrera-Viedma et al., 2011) for links and references.

Recent work has added web services to the Sci2 Tool and selected workflows can now be run online. Other efforts aim to expand the adoption of OSGi/CIShell in support of algorithm and tool plugin implementation and sharing across scientific boundaries. Tools that are OSGi/CIShell compatible comprise TEXTrend (http://textrend.org) led by George Kampis at Eötvös Loránd University, Budapest, Hungary supports natural language processing (NLP), classification/mining, and graph algorithms for the analysis of business and governmental text corpuses with an inherently temporal component and DynaNets (http://www.dynanets.org) coordinated by Peter Sloot at the University of Amsterdam for the study of evolving networks, or SISOB (http://sisob.lcc.uma.es) an observatory for science in society based in social models.

Much of the development time for the SDB for the last year has been focused on adding data to the system and refactoring code to make it easier to manage and update. Going forward, we plan to implement an API to further ease access and usage of the SDB and we are exploring an RDF conversion to add SDB to the Web of Linked Open Data (Heath & Bizer, 2011). In addition, we are considering a visual interface to SDB that uses Sci2 Web services to empower users to interactively explore, analyze, and visualize search results.

Documentation and teaching of tool functionality and workflows are important for research and practice. SDB and Sci2 are used in the Information Visualization MOOC (http://ivmooc.cns.iu.edu) which debuted in Spring 2013 to over 1,700 users, making existing and new workflows available via video tutorials to a much broader audience.

Acknowledgements

The Scholarly Database is funded by the National Science Foundation under Grant No. IIS-0238261. SDB and Sci2 are supported by the National Science

Foundation under Grants No. IIS-0513650, SBE-0738111, a James S. McDonnell Foundation grant, and the Cyberinfrastructure for Network Science center at the School of Library and Information Science at Indiana University. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

References

Alencar, A. (2010). CIShell: Extract Directed Network Retrieved January 24, 2013, from http://wiki.cns.iu.edu/display/CISHELL/Extract+Directed+Network

Belter, C. (2012). Visualizing Networks of Scientific Research. *Information Today, Inc., 36*(3). Retrieved from http://www.infotoday.com/online/may12/Belter-Visualizing-Networks-of-Scientific-Research.shtml

Biberstine, J. R. (2011). CIShell: Proportional Symbol Map Retrieved January 24, 2013, from http://wiki.cns.iu.edu/display/CISHELL/Map+of+Science+via+Journals

Biberstine, J. R. (2012). CIShell: Proportional Symbol Map Retrieved January 24, 2013, from http://wiki.cns.iu.edu/display/CISHELL/Proportional+Symbol+Map

Börner, K. (2011). Plug-and-Play Macroscopes. *Communications of the ACM 54*(3), 60-69.

Börner, K., Klavans, R., Patek, M., Zoss, A., Biberstine, J. R., Light, R., Boyack, K. W. (2012). Design and Update of a Classification System: The UCSD Map of Science. *PLoS ONE, 7*(7), e39464. doi: doi:10.1371/journal.pone.0039464

Cobo, M. J., López-Herrera, A. G., Herrera-Viedma, E., & Herrera, F. (2011). Science mapping software tools: Review, analysis, and cooperative study among tools. *Journal of the American Society for Information Science and Technology, 62*(7), 1382-1402.

Heath, T., & Bizer, C. (2011). *Linked Data: Evolving the Web Into a Global Data Space*. San Rafael, CA: Morgan & Claypool Publishers.

Johnson, D. W., & Jankun-Kelly, T. J. (2008). *A scalability study of web-native information visualization*. Paper presented at the Graphics Interface Conference 2008, Windsor, Ontario, Canada.

Kleinberg, J. (2003). Bursty and hierarchical structure in streams. *Data Mining and Knowledge Discovery, 7*(4), 373-397.

Kosecki, S., Shoemaker, R., & Baer, C. K. (2001). Scope, characteristics, and use of the U.S. Department of Agriculture intramural research. *Scientometrics, 88*(3), 707-728.

LaRowe, G., Ambre, S., Burgoon, J., Ke, W., & Börner, K. (2009). The scholarly database and its utility for Scientometrics research. *Scientometrics, 79*(2), 219-234.

Price, D. J. d. S. (1963). *Little Science, Big Science*. New York: Columbia University Press.
Robertson, G., Ebert, D., Eick, S., Keim, D., & Joy, K. (2009). Scale and complexity in visual analytics. *Information Visualization, 8*(4), 247-253. doi: 10.1057/ivs.2009.23
Suda, B. (2012). The top 20 data visualization tools. *.net*. Retrieved from http://www.netmagazine.com/features/top-20-data-visualisation-tools

OPTIMIZING RESEARCH IMPACT BY ALLOCATING FUNDING TO RESEARCHER GRANT PORTFOLIOS: SOME EVIDENCE ON A POLICY OPTION (RIP)

Rigby, John [1]*, and Julian, Keith[1]

*Corresponding Author
Manchester Institute of Innovation Research, Manchester Business School, University of Manchester, Oxford Road, United Kingdom, M13 9PL

Abstract
The attempt to maintain and indeed increase the quality of funded research in an era of huge pressure on the budgets of funding bodies has led, perhaps not surprisingly, to urgent discussion of how best to use the available financial resources. Despite the fact that funding bodies usually require undertakings from grant applicants not to seek "double funding", concern remains that duplication of funding may still occur. Moreover, the argument has been made that funding bodies should take far more account of the whole researcher portfolio when resources are allocated or even when researchers apply for grants. But attempts at top-down management of research grant allocations by funding bodies raise difficult questions. Who is in the best position to implement attempts to target research funding precisely to research topics? Some researchers might well seek duplicate funds for their work but can funding bodies singly or jointly handle the issue of directing research resources under conditions of uncertainty more ably than researchers? Analysis of the papers funded by two major grant awarding bodies that each support research in the area of molecular biology suggests that funding by both of these organisations on a per paper basis [our proxy for a discrete knowledge generating activity] leads to research with higher impact. We believe that funding from more than one source may in some circumstances lead to positive interaction rather than waste of resources.

Conference Topic
Collaboration Studies and Network Analysis (Topic 6); Modeling the Science System, Science Dynamics and Complex System Science (Topic 11)

Introduction
An issue of perennial interest in science policy is the contribution made by funding organisations through their peer review systems and their grants to the progress of scientific enquiry. While the literature indicates a degree of consensus that funding is essential to research, there remains disagreement about the role of funding bodies' peer review systems as guarantors of or influencers of the novelty and impact and the extent of any interaction of funding provision between them (Baldi 1998; Rigby 2012). The provision of grant schemes has, over a long period, diversified to meet a range of needs of different types of researchers at

different stages of their careers. However, there remains concern that peer review systems, even of the most respected grant awarding bodies, protect the status quo (Nicholson and Ioannidis 2012).

There have been a number of recent suggestions in the research policy literature of ways to change the allocation of resources so as to ensure greater impact (Heinze, Shapira et al. 2009; Luukkonen 2012). An earlier paper by Heinze (Heinze 2008) has proposed a number of changes to the processes by which research grants are awarded to bypass the status quo, including the creation of new agencies and the support of research proposals which are ”high risk”.

However, in addition to using the ”uncertainty” criterion or the perceived risk of a research proposal, assessed by the extent of disagreement between the peer reviews of a proposal, Heinze also suggests that grant awarding bodies should make their awards subject to two principles the first of which is uncontroversial, i.e. that the grant application is assessed on the basis of the scientific case proposed. But his second principle is however quite radical. It proposes that any funding should be awarded in light of the funding already allocated to the researchers whose proposal is under review: ”Hence, funding agencies should consider better aligning their resources with the existing funding of successful applicants” (Heinze 2008 page 315).

The belief that greater ”top-down coordination” of research funding will prevent duplication and achieve greater scientific impact is beginning to receive more weight, partly in light of the discovery of instances of apparent duplication of research funding (”double funding”), reported in 2012 (Reich 2012) and a more recent Nature editorial (Reich and Myhrvold 2013) based on the research of Garner, et al (Garner, McIver et al. 2013). An interesting irony here is that some years ago, Lewison and Dawson (1998) proposed that the count of funding acknowledgements generally predicted impact, a desirable outcome, arguing that research (papers) resulting from such funding would have had more peer review. However, the suggestion is now emerging that funding from multiple agencies (which would lead to multiple funding acknowledgements on papers) might be a sign of duplication of resources, a state of affairs which should be avoided.

The authors of these recent studies on duplication are however careful to acknowledge that their claims of duplication have been limited by the data they have been able to use, and they believe that only when full grant applications, grant summaries and resulting publications are available will it be possible to assess the true extent of duplication of funding for research activities. Nevertheless, this recent evidence suggests there is now a need to examine closely the rules that exist to ensure fair allocation of research funds to research work, including those procedures that exist to re-allocate research money in the event of changes to the work being carried out.

However, the attempt to optimize scientific outcomes by creating further rules and processes than already exist to prevent double funding ("double dipping") and to impose greater oversight of the research priority setting of scientists by administrators is not in our view likely to be easy to implement and may not be realistic policy. As this research in progress paper argues, bibliometric evidence does show that the support of certain funding bodies may increase the level of impact achieved rather than entailing duplication and waste of resources. Using the data on papers funded by two organisations that operate within the same areas of science (molecular biology) and which attempt to achieve similar ground breaking research within this research field, we show that papers with funding acknowledgements from both the European Molecular Biology Organisation and the Human Frontier Science Program have greater impact than papers with just one or other of these organisation's funding acknowledgements.

Methods

Meta-data on papers from the Web of Knowledge published in the period 2008-2012 which had funding acknowledgements from either the European Molecular Biology Organisation or the Human Frontier Science Programme, or from both of these organisations, were downloaded and analysed. Papers were chosen from 2008 as this is the first year in which there is full coverage of funding acknowledgement data within the Thomson Reuters Web of Knowledge citation index.

Papers were uploaded to VantagePoint and then cleaned to disambiguate funding acknowledgement and author address data as this is not standardized and many different occurrences of funding body and addresses names are present. A threefold categorization of papers was created: EMBO funding acknowledged, HFSPO funding acknowleged, funded by both EMBO and HFSP. In all 21903 funding acknowledgements were recorded across the whole data set with 5029 acknowledgements (23%) from these two organisations. As EMBO operates a significant number of research laboratories, papers that had been written by authors whose addresses indicated an EMBO laboratory were considered as EMBO papers for the purpose of the analysis. Articles and Review papers were treated separately as their funding is *a priori* different in origin.

The papers were compared at the level of the paper, as few papers had more than one grant number applying to them. Papers were then compared in terms of their citedness, the count of authors, the count of author affiliations, and the count of funding acknowledgements. Articles were analysed separately from reviews. Book chapters and other forms of publication of which there were small numbers in the data set were not included in the analysis.

Results

The Kruskal-Wallis ranks test was used to determine differences between the three groups of papers. The Kruskal-Wallis is a non-parametric test of ranks, using an overall difference in ranks. It is a suitable test where samples are of different sizes and distributions of variables cannot be assumed to be normal. P-values less than 0.05 imply a statistically significant difference amongst the medians at the 95.0% confidence level. The results of the four sets of tests are given below. The table shown below displays the results of the Kruskal-Wallis test (ranks, test statistic and p-value) for each of the four comparisons.

Table 1. Kruskal-Wallis Tests of Group Differences: Citations Per Paper; Authors Per Paper; Author Affiliations Per Paper; Funding Acknowledgements Per Paper

Kruskal-Wallis Test for Total Citations	**Articles**		**Review Papers**	
Group	Sample Size	Average Rank	Sample Size	Average Rank
EMBO and HFSP	245	2463.31	30	268.63
EMBO	1550	2148.11	218	271.81
HFSP	2412	2039.16	295	272.4
Test Statistic	30.62		0.016	
P Value	2.23E-07		0.991	
Kruskal-Wallis Test for Total Authors	**Articles**		**Review Papers**	
Group	Sample Size	Average Rank	Sample Size	Average Rank
EMBO and HFSP	245	2410.73	30	270.567
EMBO	1550	2345.29	218	290.782
HFSP	2412	1917.78	295	258.266
Test Statistic	134.844		5.91264	
P Value	0.000		0.05201	
Kruskal-Wallis Test for Total Author Affiliations (Count of Institutions)	**Articles**		**Review Papers**	
Group	Sample Size	Average Rank	Sample Size	Average Rank
EMBO and HFSP	245	2292.96	30	294.483
EMBO	1550	2247.61	218	277.048
HFSP	2412	1992.52	295	265.983
Test Statistic	50.2912		1.45241	
P Value	1.20E-11		0.483741	
Kruskal-Wallis Test for Count of Funding Acknowledgements	**Articles**		**Review Papers**	
Group	Sample Size	Average Rank	Sample Size	Average Rank
EMBO and HFSP	245	2908.5	30	408.783
EMBO	1550	2261.37	218	281.557
HFSP	2412	1921.15	295	251.027
Test Statistic	192.342		29.7076	
P Value	0		3.54E-07	

Articles (but not review papers) that acknowledged funding from both sources were statistically more likely to have more citations than papers funded separately, more likely to have more institutions involved and more likely to have more authors involved in their production. The greater likelihood of jointly funded papers (EMBO and HFSP) having more funding acknowledgements might be related to their having *a priori* one more funding acknowledgement than a paper funded by only one of the organisations (i.e. either EMBO or HFSP) but the effect of this on distributions is work in progress.

Discussion

The results of the analysis conducted here indicate that where researchers receive funding from EMBO and from HFSP at the same time, rather than separately, (i.e. from either one or the other of the two organisations), the resulting papers achieve a higher citation count. This may be attributable in part to the presence of more organisations and authors in jointly funded papers. The analysis does not provide hard evidence of a net benefit of ”simultaneous funding”. To achieve this, there would need to be control of type of research undertaken to reduce the risk that the different categories of funding have chosen different forms of research projects with potentially different levels of impact.

This research in progress paper has sought to encourage discussion on the issue of whether research funding bodies might take account of the funding that researchers are already receiving or about to receive in order to reduce duplication of funding and to enhance research impact. Steps to take control of, manage, or take account of researcher funding portfolios may appear desirable in an era of increasing austerity as this might promise to reduce redundancy of funding, particularly where certain funding bodies are known to support research in the same or very similar areas. However, the attempt to allocate funding by grant awarding bodies to remove or reduce the risk of redunancy and to optimize research output raises serious questions about effectiveness of outcome, as well as about the confidentiality of the work of researchers.

Policy researchers should investigate this area in more detail as the evdience suggests that simultaneous funding may be beneficial, and that is leads to the provision of a level of resources for researchers that gives achieve greater impact on any given research problem. Funding organisations and researchers should between them decide on the basis of further examination of this question where responsibility should lie for the allocation of resources to research problems. It is in our view premature to assume that funding bodies should control the process of resource allocation as it researchers which are more likely to know best how to match funding to the specific challenges of the work they are undertaking.

References

Baldi, S. (1998). "Normative versus social constructivist processes in the allocation of citations: A network-analytic model." American Sociological Review **63**(6): 829-846.

Garner, H. R., L. J. McIver, et al. (2013). "Research funding: Same work, twice the money?" Nature **493**(7434): 599-601.

Heinze, T. (2008). "How to sponsor ground-breaking research: A comparison of funding schemes." Science and Public Policy **35**(5): 302-318.

Heinze, T., P. Shapira, et al. (2009). "Organizational and institutional influences on creativity in scientific research." Research Policy **38**(4): 610-623.

Lewison, G. and G. Dawson (1998). "The effect of funding on the outputs of biomedical research." Scientometrics **41**(1-2): 17-27.

Luukkonen, T. (2012). "Conservatism and risk-taking in peer review: Emerging ERC practices." Research Evaluation **21**(1): 48-60.

Nicholson, J. M. and J. P. A. Ioannidis (2012). "Research grants: Conform and be funded." Nature **492**(7427): 34-36.

Reich, E., S. and C. Myhrvold, L. (2013). "Funding agencies urged to check for duplicate grants: Nature probe reveals lack of oversight of researchers who win two grants for similar projects." Nature **493**

Reich, E. S. (2012). "Duplicate-grant case puts funders under pressure." Nature **482**(7384): 146-146.

Rigby, J. (2012). "Looking for the impact of peer review: does count of funding acknowledgements really predict research impact?" Scientometrics: 1-17.

PATENTS IN NANOTECHNOLOGY: AN ANALYSIS USING MACRO-INDICATORS AND FORECASTING CURVES

Douglas H. Milanez[1], Leandro I. L. Faria[2], Roniberto M. do Amaral[3] and José A. R. Gregolin[4]

[1] *douglas@nit.ufscar.br*
Federal University of Sao Carlos, Center for Information Technology in Materials, PhD student from Graduate Program in Materials Science and Engineering , Washington Luis Highway, km 235, São Carlos – SP (Brazil)

[2] *leandro@nit.ufscar.br*
Federal University of Sao Carlos, Center for Information Technology in Materials, professor from the Dept of Information Science, Washington Luis Highway, km 235, São Carlos – SP (Brazil)

[3] *roniberto@nit.ufscar.br*
Federal University of Sao Carlos, Center for Information Technology in Materials, professor from the Dept of Information Science, Washington Luis Highway, km 235, São Carlos – SP (Brazil)

[4] *gregolin@nit.ufscar.br*
Federal University of Sao Carlos, Center for Information Technology in Materials, professor from the Dept of Materials Science, Washington Luis Highway, km 235, São Carlos – SP (Brazil)

Abstract

The aim of this study is to analyze the technological output in nanotechnology using patent indicators, which were developed based on a set of 189,481 records recovered from the Derwent Innovation Index database. Future trends of worldwide nanopatenting were evaluated using logistic growth curves and annual growth rates from 1995 to 2010, while the patenting activity from the main countries and technological domain and subdomain were assessed from the perspective of worldwide, USPTO, TRIAD and TETRAD patent documents from 2000 to 2010. Outcomes of nanotechnology patent activity have generated interesting discussions as they suggest that technological development has reached a maturation stage apparently. Although China´s share of patents is small in some cases, it was the only country to constantly increase the number of patents from a worldwide perspective. In contrast, the USA and the EU were the most active in the USPTO, TRIAD and TETRAD cases, followed by Japan and Korea. The technological subdomains of main interest from countries/region changed according to the perspective adopted, although there was a clear bias towards Semiconductors, Surface Treatments, Electrical Components, Macromolecular Chemistry, Materials-Metallurgy, Pharmacy-Cosmetics and Analysis-Measurement-Control subdomains. Finally, monitoring

nanotechnology advances should be constantly reviewed in order to confirm the evidence observed.

Conference Topic

Technology and Innovation Including Patent Analysis (Topic 5)

Introduction

Recently, there has been considerable interest in evaluating nanotechnology research developments due to its high potential to promote significant innovation in products, processes, materials and devices and to benefit the society (Milanez, 2011; Salerno, Landoni, & Verganti, 2008). This has encouraged worldwide scientific efforts and governmental programs to fund research in nanotechnology, especially from the United States, Japan, China, Korea and the European Community. For instance, global public spending has increased from approximately US$ 4.5 billion to US$ 10 billion between 2005 and 2010 (Observatorynano, 2012; Roco, 2005). Only the US National Nanotechnology Initiative 2012 budget was estimated at US$ 1.7 billion (United States, 2012). The current advances in nanotechnology have been quite striking and evident from publications and patent document data (Dang, Zhang, Fan, Chen, & Roco, 2010; Kostoff, Koytcheff, & Lau, 2007; Porter, Youtie, Shapira, & Schoeneck, 2008)

To follow advances in nanotechnology, it is preferable to use quantitative methodologies as they are objective and are able to compare results. However, monitoring and evaluating nanotechnology developments using bibliometric approaches is a challenge due to its initial stage of development and interdisciplinarity, which can make retrieving information and establishing knowledge boundaries difficult. (Milanez, 2011; Porter et al., 2008; Salerno et al., 2008). For instance, nano-related patent classification codes emerged from the United States Patent and Trademark Office (USPTO), the European Patent Office (EPO) and the World Intellectual Property Organization (WIPO) as an effort to attempt to precisely describe and monitor this area (Scheu et al., 2006), whereas they alone are not sufficient to retrieve all information (Porter et al., 2008). Furthermore, there is still a lack of specific regulatory frameworks describing safety procedures accurately, environmental and health effects and risks of manipulating nanomaterials (Salerno et al., 2008), which make future event-changes uncertain. Therefore, new research to monitor its development can minimize the gaps in knowledge and future uncertainties.

Bibliometric analysis of patent documents is an approach to evaluate trends in technological developments, as well as country and competitors interests. An important indicator is the annual number of patents in nanotechnology, but few researchers have forecasted the future developments of this area, for example, using the logistic growth curve method. According to Martino (1993), the logistic curve is an extrapolation method that assumes the past of a time series contains all

the information needed to forecast the future growth of a time series up to a limit established. Cheng and Chen (2008) applied this method to USPTO patent data of nanosized ceramic powders from 1970 to 2005 and they observed that ceramic nanomaterials were in the initial growth periods of a technological life cycle. Milanez, Amaral, Faria & Gregolin (2013) also forecasted nanocellulose scientific and technological activities using the logistic growth curve and observed emerging stage of nanocellulose. Even though Alencar, Porter and Antunes (2007) observed this logistic growth trend in nanotechnology data, they did not predict its future development and life cycle stage. Besides the development stage, the logistic growth curve may also support planning and future investment (Martino, 1993).

Another issue addressed in papers is to verify which countries or regions were leading or competing in the "nanorace". Some studies only considered the USPTO or the EPO patents as a reference due to their economical importance (Chen, Roco, Li, & Lin, 2008; Cheng & Chen, 2008; Z. Huang et al., 2003; Z. Huang, Chen, Chen, & Roco, 2004; Hullmann & Meyer, 2003; Igami, 2008), but limiting these repositories make it difficult to comparatively analyse worldwide because of emerging developing economies, such as China, Russia, Brazil and India (Glänzel, Debackere, & Meyer, 2007). Other studies considered data from worldwide databases, such as *Derwent Innovations Index* (Alencar et al., 2007; Wang & Guan, 2012) and *Espacenet* (Dang et al., 2010). Dang et al. (2010) included an assessment of the patent application published at the main patent offices, yet their analysis did not consider the economic value of patent filing, such as those from the *tridiac patent families* (also known as TRIAD). According to the *OECD Patent Statistic Manual* (2009), the TRIAD is a set of patent applications filed at the Japanese Patent Office (JPO), EPO and USPTO (in this last case, patents filed after 2001 as the USPTO did not publish patent files before 2000). TRIAD statistical analysis has an advantage to improve international comparability and include patent families that are typically of high value as it is assumed that the additional cost of protection in different countries is worthwhile (OECD, 2009). Recently, Wang and Guan (2012) evaluated the worldwide trend of patenting in nanotechnology using a dataset from *Derwent Innovations Index*. They included the TRIAD patents and observed a strong presence of the United States, the European Union and Japan. By contrast, due to the fact that China is challenging the leading science and technology countries and the Chinese market has become one of the most important in the world, analysis should include the TETRAD patent families (Glänzel et al., 2007)

Evaluating the technological sector performance of nanotechnologies and their involvement with countries is another task of a wide range of studies. It is common to measure this key issue by patent classification information, such as US Patent Classification (Z. Huang et al., 2003, 2004) and International Patent Classification (IPC) (Alencar et al., 2007; Dang et al., 2010). However, the

general purpose of patent classifications is to describe the content of the patent document in detail, including issues about the invention, functionality and application. Consequently, they are too disaggregated to analyse technological tendencies because different classifications may be related to the same technological domain. Moreover, a single patent document may contain many classification codes and belong to one or more sector. In order to overcome this gap, patent documents can be aggregated according to their classification code and improve the analysis. Wang et al. (2012) classified the patent documents in five industrial areas and 35 fields of applications. They also observed a concentration of nanopatenting, mainly in the Chemical industrial area, followed by Instruments and Electrical Engineering areas. Their outcome also showed countries' fields of application highlighting China in Materials and Metallurgy, the United States in Pharmaceuticals and Semiconductors, Japan in Macromolecular Chemistry and Optics, Korea in Semiconductors and the European Union in Pharmaceuticals. Nevertheless, they have used only the first IPC in the aggregation process, thus relevant information was lost and probably could have benefitted from a specific industrial area or field of application.

The aim of this paper is to update and pursue studies analyzing technological development in nanotechnology using bibliometric indicators based on patent documents. Development in worldwide patenting and future trends was evaluated using logistic growth curves and the annual growth rate, while the behavior of the main countries/region and the technological domain and subdomain were assessed by the share of patents. Analyses were also carried out in four aspects of patenting: considering the worldwide and patent documents from the USPTO, TRIAD and TETRAD.

Method and procedures adopted

Procedures for collecting and analyzing the patent data

Technological indicators were developed using the patent data indexed in the *Derwent Innovations Index* (DII) (Thomson – ISI, US). DII has the advantage of covering patent bibliographic information from worldwide main repositories, allowing for searches in multiple bibliographic fields, such as the title, abstract, inventors, assignees and International Patent Classification (IPC). It also makes use of complex Boolean search expressions. Furthermore, DII records are aggregated according to their family patent[124], which provides an analysis of different contexts without duplicating the document.

Endnotes:

[124] A patent family is a core of published patent documents referring to the same invention and applied in different countries by way of the priority or priorities of a particular patent document. A patent document can be referred to as an applied or a granted patent.

Nanotechnology patent documents were retrieved by using the modularized Boolean search strategy suggested by Porter et al. (2008). Defining a search expression for nanotechnology is a challenge due to its interdisciplinary nature. However, Huang, Notten and Rasters (2011) comparatively reviewed various search strategies and they concluded that most of the final rankings are similar due to the fact that these search strategies share a group of key words and terms, although the quantity of documents retrieved are different. Moreover, the modular search strategy chosen has some advantages. First, it combines a number of nano-related terms from earlier studies and others accurately selected by experts. Secondly, it can be used in multiple databases retrieving a large-scale core of relevant data (Porter et al., 2008; Wang & Guan, 2012). The search was carried out on 23rdJanuary, 2013 and 189,481 records were retrieved considering the time span until 2012. All data were collected and imported to the bibliometric software *VantagePoint* (version 7.0, Search Technology Inc, US). A set of bibliometric indicators were developed:

- Number of patents per year and annual growth rate from 1995 to 2010;
- Accumulated number of patent documents per year from 1995 to 2010;
- Forecasting growth curves based on the accumulated patent documents;
- Share of patent documents by the main patenting country/region[125] from 2000 to 2010;
- Share of patent documents by technological domains and subdomains from 2000 to 2010;
- Share of patents by the main patenting countries/region according to their most relevant technological subdomain from 2000 to 2010.

The period of analyses was limited up to 2010 because of the delay between an application and publication of a patent document, regularly 18 months in most countries (MOGEE, 1997). To evaluate the country performance and determine the year when a patent document was first published, the earliest priority of DII records was selected in view of the fact that they do not provide the nationality of inventors or applicants. It was assumed that the country and the year from the earliest priority refer to the place and the period the invention was developed. In order to analyze the main technological subdomains related to nanotechnology patenting, the data were processed according to the subdomains´ classification suggested by the *Observatorie dês Sciences et dês Techniques* (OST, 2010). In the process, patent documents were grouped into seven technological domain and thirty technological subdomains according to their IPC code.

[125] The European Union patent documents included patents from Austria, Belgium, Bulgaria, Czech Republic, Cyprus, Denmark, Estonia, France, Finland, Greece, Germany, Hungary, Ireland, Italy, Latvia, Lithuania, Luxemburg, Malta, Netherlands, Poland, Portugal, Romania, Slovenia, Slovakia, Spain, Sweden, United Kingdom, and European Patent Office.

Four different cases were considered in order to assess nanotechnology patenting: patent documents from worldwide repositories, and from USPTO[126], TRIAD[127], and TETRAD[128].

Annual growth rate and share of patent calculations

The annual growth rate (G_i) was calculated using equation 1, where N_i is the number of patent documents in year "i" and N_{i-1} is the number of patent documents in year "i-1".

$$G_i = \frac{(N_i - N_{i-1}) * 100}{N_{i-1}} \quad (1)$$

The percentage share (S) of patent documents was calculated using Equation 2, where S_i is the number of patent documents from a country/region or a technological subdomain (i) and S_t is the total of patent documents in the context of the analysis (t).

$$S = \frac{S_i * 100}{S_t} \quad (2)$$

Logistic Growth Curves Calculation

The increase in the number of accumulated patents in nanotechnology was predicted using the logistic growth curve which is calculated according to Equation 3. *L* is the upper limit of the growth of variable *y*, *t* is time, *a* and *b* are coefficients obtained by fitting the growth curve to the known data, and *e* is the base of natural logarithms (Martino, 1993).

$$y = \frac{L}{1 + ae^{-bt}} \quad (3)$$

Three upper limits (*L*) were tested properly to state the future development of nanotechnology patent documents. These upper limits were chosen considering their best fit to the real annual cumulative data from 1995 to 2010. Furthermore, inflection points of the three curves were obtained in order to delimit the emerging and maturity stages (Martino, 1993; Cheng & Chen, 2008).

Results and discussion

General and future prospects on nanopatenting

Worldwide patenting boomed from 1995 to 2010, as can be seen in Figure 1. In this period, the number of patents grew 1.447%, from 1.451 documents in 1995 to

[126] Records with at least one US patent number were considered as files at USPTO.
[127] For TRIAD patents, records with at least one US patent number, one EPO patent number and one JPO patent number were considered.
[128] For TETRAD patents, records with at least one US patent number, one EPO patent number, one JPO patent number, and one SIPO (State Intellectual Property Office of the People's Republic of China) patent number were considered.

22.442 in 2010. The annual growth rates followed an interesting trend, if the value from 2001 is ignored: the growth rate rose rapidly until 1998 and then gradually declined until 2010. Regarding the 2001 growth rate value, although this is the year after launching the US nanotechnology program (the *National Nanotechnology Initiative*), which is one of the most important worldwide programs encouraging other countries' programs, a detailed analysis of the data showed that the peak is a consequence of non-regular behavior from Chinese patenting (see Figure 3). Some facts may have influenced this result from China, such as the impulse of Chinese research in nanotechnology from two research programs in 1999 (the *National Key Basic Research Program* and the *Applied Research on Nanomaterials Program*) (Galembeck & Rippel, 2006; Milanez, 2011). Another fact was the restructuring of their intellectual property system to become a member of the Agreement on Trade Related Aspects of Intellectual Property Rights (TRIPS), one of the essential requirements for joining the World Trade Organization in 2000. The same sudden increase in the number of patents from China in 2001 was discussed by Hu and Jefferson (2009).

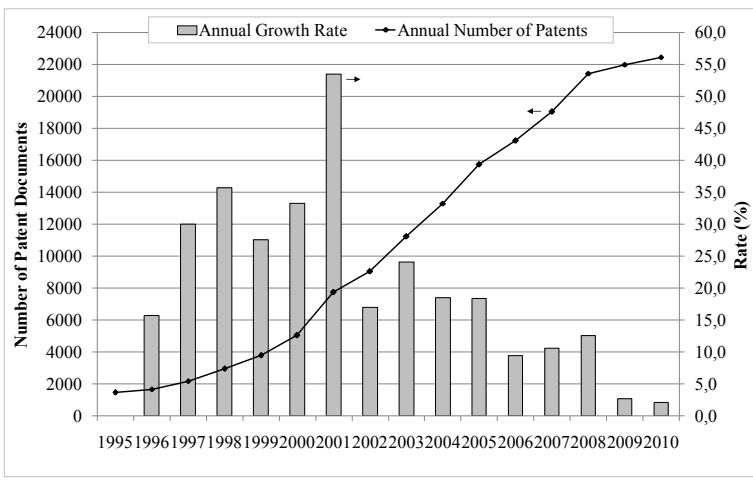


Figure 1. Annual number and annual growth rate for patent documents in nanotechnology from 1995 to 2010.

Interestingly, the growth rates from 2009 and 2010 suggest that the annual number of patents will grow slowly in coming years. The accumulated number of patent documents also shows a gradual decrease towards logistic growth pattern in coming years, as can be seen in Figure 2. The growth curves indicated the inflection point, which occurred between 2008 and 2009, indicating the beginning of deceleration in nanotechnology patenting. According to Martino (1993) and Cheng et al (2008), this slowing down might be interpreted as the beginning of maturation from the technological development and new growth cycles may occur in the future due to scientific advances. Although the patenting activity might

have experienced effects from the recent financial crisis and ongoing economic uncertainty, these outcomes generated a rather interesting hypothesis and considerations to be discussed. First of all, the inflection points were quite near to the present moment, thus no additional data was readily available to confirm this trend. Secondly, as nanotechnology is considered as an emerging interdisciplinary field (Milanez, 2011; Salerno et al, 2008; Porter et al, 2008), new nano-related terms might have appeared overtime. Therefore the data used to develop these indicators could be incomplete. In this case, Arora, Porter, Youtie & Shapira (2012) have updated the nanotechnology modular search strategy used in this study (from Porter et al., 2008) with 33 new terms, which might change the overall picture in Figures 1 and 2. Moreover, due to the fact that nanotechnology can be found in various subfields and each one has their own way of working, the forecast of its current state could have been made applying the growth curve model to the subfields separately and then the combinations of these results could depict an overall picture. Other papers (Cheng & Chen, 2008; Milanez et al., 2013) applied the growth curve model to nanocellulose and nanosized ceramics and their findings suggested that these nanomaterials were in their emerging period (before the inflection point). This result corroborates with the hypothetical need to forecast the nanotechnology subfields and nanotechnology as a whole.

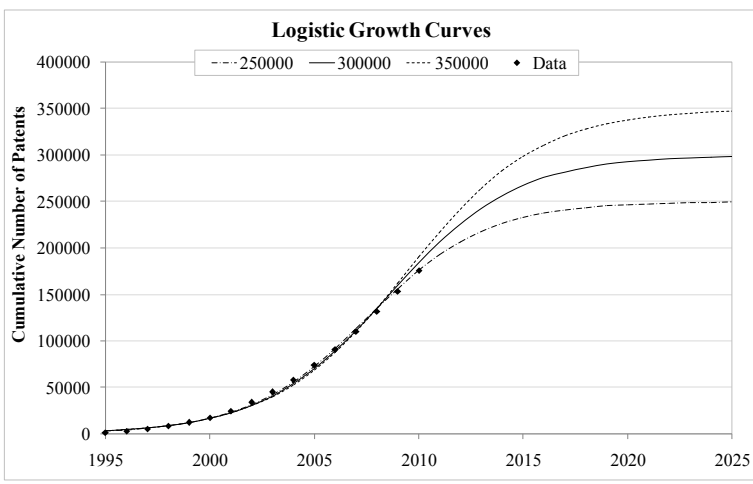


Figure 2. Accumulated number of patent documents from 1995 to 2010 and forecasting growth curves for nanotechnology.

Nanotechnology is a multidisciplinary field and seeing it as only one area in the future is complicated. What could be done is to foresee other subfields and combine them to have a general overview of the development stage of nanotechnology. Finally, another issue concerns the close link between scientific advances and technological development in nanotechnology. Scientific

knowledge strongly emerged from 2000 to 2011[129], but our outcomes (from Figure 1 and 2) suggest they do not lead to technological development, probably due to the challenges regarding industrial scale production and risks of nanomaterials to human health and the environment (Salerno et al, 2008).

The dynamic of patenting for the main countries/region

The United States (US), Japan, Korea, the European Union (EU) and China shared 93.5% of the worldwide patents in nanotechnology from 2000 to 2010, as can be seen from Table 2. This result corroborates with others from the literature (Dang et al., 2010; Glänzel et al., 2007; Milanez, 2011; Wang et al., 2012), although some differences may occur due to the search strategy used. China has already accumulated more patent documents than the USA, sharing 27.0% of the whole documents retrieved. According to worldwide trends in Figure 3, China was the only country that showed a steady growth in the period analyzed, probably due to its economical situation. On the other hand, although the USA, the EU, Japan and Korea have large quantities of patent documents, they have declined in recent years in the period analyzed.

The most striking result to emerge from Table 2 is the small number of patent documents that China had in the USPTO, the TRIAD and the TETRAD, which raises an issue as to whether their nanotechnologies are potential economically or not. In spite of the logical fact that they are leaders in patenting in USPTO, the USA was also in first place at TRIAD and TETRAD (45.0% and 44.3%, respectively) followed by the EU (31.3% and 30.9%), Japan (15.1% and 15.4%) and Korea (4.78% and 5.20%). This result shows evidence of the economic value and their interest to protect their developments in other markets.

Table 2. Share of nanotechnology patent documents from 2000 to 2010 among the main countries/region at the worldwide, USPTO, TRIAD and TETRAD cases.

Country/ Region	*Worldwide Share (%)*	*USPTO Share (%)*	*TRIAD Share (%)*	*TETRAD Share (%)*
US	26.8	61.9	45.0	44.3
EU	11.3	14.3	31.3	30.9
Japan	17.0	10.4	15.1	15.4
Korea	11.4	6.82	4.78	5.20
China	27.0	2.13	0.76	1.19

Figure 3 also provides the annual number of patent documents of the USPTO, TRIAD and TETRAD and states the patenting activity of the main

[129] We performed a quick analysis from scientific publications indexed in the Science Citation Index Expanded and Social Science Citation Index (Web of Science) from 2000 to 2010 using the modular search strategy for nanotechnology (Porter et al., 2008). On average, scientific publications went up by 13.09% in the period considered and the annual growth rates of 2008, 2009, 2010 and 2011 were 15.57%, 7.79%, 8.93% and 13.81%, respectively.

countries/region in the period considered. Once more, the US dominated the annual number followed by the EU, Japan, Korea and China and, in general, the number of patent documents increased up to a specific year for each country/region and then decreased. These trends are similar to the ones observed worldwide, except for China. However, there was a sharp fall in 2009 and 2010 and this could partially be a consequence of delaying processes of indexing due to the lack of patent data because of the Patent Cooperation Treaty (PCT) process or the search expression used. Moreover, the effect of the recent financial crisis would have affected their patent activity.

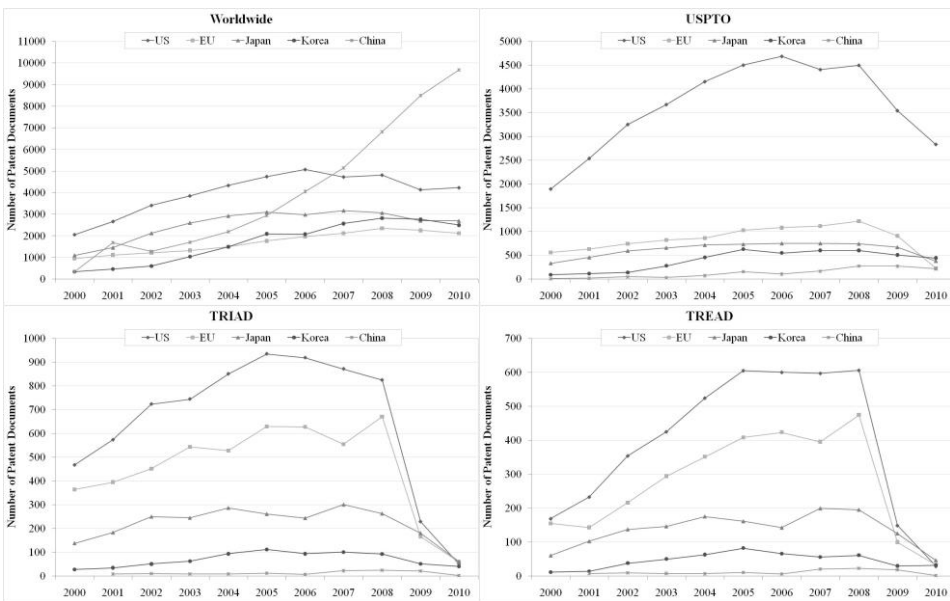


Figure 3. Worldwide, USPTO, TRIAD and TETRAD annual number of patent documents for the main countries/region.

The dynamic of patenting for technological domains and subdomains

Table 3 compares the shares of patent documents according to technological domains and subdomains in worldwide, USPTO, TRIAD or TETRAD situations.

Table 3. Share of patent documents according to technological domains and subdomains in worldwide, USPTO, TRIAD or TETRAD situations.

Technological ***Domain***/Subdomain	*Worldwide Share (%)*	*USPTO Share (%)*	*TRIAD Share (%)*	*TETRAD Share (%)*
Electronic-Electricity	31.0	40.5	37.4	39.0
Electrical Components	14.2	16.7	19.6	21.3
Audiovisual	2.11	3.29	3.22	3.47
Telecommunications	0.99	1.57	1.39	1.38
Data Processing	2.07	3.77	3.44	3.53

Semiconductors	16.8	24.0	22.2	22.7
Instrumentation	24.6	32.3	37.7	32.1
Optics	8.28	11.1	12.3	11.9
Analysis-Measurement-Control	11.6	15.8	18.8	14.4
Medical Engineering	5.35	6.82	9.57	7.68
Nuclear Techniques	1.03	1.83	2.17	1.82
Chemistry-Materials	44.0	46.1	60.7	63.9
Organic Chemistry	3.98	6.81	11.8	11.0
Macromolecular Chemistry	13.0	13.3	22.4	25.1
Basic Chemistry	8.30	9.93	17.0	18.6
Surface Treatments	13.5	21.7	25.1	26.6
Materials-Metallurgy	16.5	13.4	20.2	22.2
Pharmacy-Biotechnology	15.8	21.7	30.8	26.7
Biotechnology	7.04	12.0	15.6	12.4
Pharmacy-Cosmetics	10.2	13.7	22.0	19.8
Agricultural and Food Products	1.12	0.90	1.62	1.73
Industrial Processes	24.1	25.1	35.5	38.0
Technical Processes	12.7	13.4	19.1	19.9
Graphical Maintenance	1.73	2.38	3.82	3.97
Work with Materials	9.89	11.6	17.6	19.8
Environment-Pollution	2.33	1.59	1.82	2.14
Agriculture and Food Equipment	0.52	0.45	0.61	0.53
Machines-Mechanics-Transp.	4.59	5.10	6.32	6.96
Machine-Tools	1.54	2.03	2.48	2.83
Motor-Pump-Turbines	0.56	0.66	0.80	0.66
Thermal Processes	1.00	0.87	0.93	1.12
Mechanical Components	0.94	0.97	1.45	1.57
Transport	0.64	0.73	1.25	1.36
Space-Weapons	0.23	0.34	0.16	0.15
Household Consumption.-Construction	3.25	2.90	2.81	3.14
Household Consumption	2.36	2.15	2.24	2.43
Construction	0.93	0.81	0.63	0.79

All patent documents were associated with at least one technological domain and subdomain showing evidence of the interdisciplinarity of technological developments on a nano-scale. Nonetheless, there was a clear bias towards Chemistry, Electronic-Electricity, Instrumentation and Industrial Process domains and Semiconductors, Surface Treatments, Electrical Components, Macromolecular Chemistry, Materials-Metallurgy, Pharmacy-Cosmetics and Analysis-Measurement-Control subdomains in the four cases analyzed. On the other hand, few developments occurred in the Household Consumption-Construction and Machines-Mechanics-Transport domains. Furthermore, even though the share may be slightly different, no technological domain or subdomain overlapped other positions with the change of perspective. Considering the Worldwide patent share, the Semiconductors subdomain concentrated the highest number (16.8%) followed closely by Materials-Metallurgy (16.5%). Electrical

Components (14.2%), Surface Treatments (13.5%) and Macromolecular Chemistry (13.0%), which were also relevant subdomains in the period analyzed. In the case of patent documents at the USPTO, besides Semiconductors (24.0%), Surface Treatments (21.7%) and Electrical Components (16.7%), another two subdomains were highlighted as important technological contexts, Analysis-Measurement-Control (15.8%) and Pharmacy-Cosmetics (13.7%). Concerning the TRIAD and TETRAD shares, there is a similarity of the main technological subdomain. Technology for Surface Treatments (25.1% and 26.6%, respectively) presented the most important subdomain, followed by Macromolecular Chemistry (22.4 % and 25.1%) and Semiconductors (22.2% and 22.7%). Materials-Metallurgy appeared, respectively, in fifth (20.2%) and fourth (22.2%) place while Pharmacy-Cosmetics stood out from TRIAD (22.0%) and Electrical Components (21.3%) from TETRAD. Additionally, it is important to clarify that the sum of percentage shared in any situation will be more than 100% because a single patent document can contain a range of different IPC codes and belong to a different technological domain or subdomain.

Table 4. Share of country/region patent documents according to their three most relevant technological subdomains in the worldwide, USPTO, TRIAD or TETRAD[130] situations.

Country/	*Worldwide*		*USPTO*		*TRIAD*		*TETRAD*	
Region	*Subd.*	*Share (%)*	*Subd.*	*Share (%)*	*Subd.*	*Share (%)*	*Subd.*	*Share (%)*
	Se	20.3	Se	21.1	PC	25.1	ST	24.6
US	ST	19.1	ST	20.1	ST	23.2	MC	23.4
	AMC	16.7	AMC	16.6	Se	21.6	Se	23.3
	MC	17.9	ST	24.7	ST	26.0	MC	29.3
EU	ST	17.2	MC	21.7	PC	25.5	ST	28.0
	TP	17.2	PC	21.3	MC	25.3	MM	23.0
	Se	22.8	Se	35.0	ST	28.5	MM	30.8
Japan	MM	22.8	EC	25.5	MM	27.9	ST	29.5
	EC	21.5	ST	24.9	Se	26.6	EC	29.3
	Se	23.9	Se	41.4	EC	38.1	EC	39.7
Korea	EC	16.7	EC	31.6	Se	31.5	Se	30.6
	MM	11.1	ST	22.7	ST	29.0	ST	28.6
	MM	21.8	EC	31.6	TP	33.9	TP	33.9
China	TP	14.4	MM	28.6	MM	29.8	MM	30.4
	MC	13.9	ST	26.0	ST	24.0	ST	23.5

The patent share of countries/region in the four situations considered can be assessed according to the main technological subdomains that they explore, as shown in Table 4. Countries/region focus on different subdomains according to

[130] In Table 4, Semiconductors = Se; Surface Treatments = ST; Analysis-Control-Measurement = ACM; Materials-Metallurgy = MM; Technical Processes = TP; Macromolecular Chemistry = MQ; Pharmacy-Cosmetics = PC; and Electrical Components = EC.

the perspective considered, although the subdomains of Korea and China tend to be close. Semiconductors was extensively explored by the US, Japan and Korea while Electrical Components was the major target for Korea and Japan. The EU had great interest in Macromolecular Chemistry in four cases. Furthermore, the EU competed against the USA for the Pharmaceutical-Cosmetics subdomain in the case of TRIAD. It should be mentioned that China has a small number of patent documents in the USPTO, TRIAD and TETRAD cases from 2000 to 2010. However, Technical Processes was a subject of interest mainly for China and Materials-Metallurgy stands out as an important technological subdomain, although Japan was a great competitor in this last subdomain. A striking result is that all countries/region were interested in the Surface Treatments subdomain, because it appears in all the evaluated situations and all countries/region, except for the worldwide case where just the USA and EU stand out.

Conclusion

This paper investigated the trends of patenting and predicts future development based on growth rates and logistic growth curve. It has also discussed the activity of the most relevant countries/region and the main technological domains and subdomains using bibliometric indicators. In the country/region and technological domain and subdomain indicators, four perspectives of patenting were considered (the worldwide trend and patent documents from USPTO, TRIAD and TETRAD) in order to assess the economic value of patent filing in nanotechnology. The following conclusions can be drawn from the present study. The share of patents indicated a clear bias towards Chemistry, Electronic-Electricity, Instrumentation and Industrial Process domains from 2000 to 2010, regardless of the perspective adopted. Considering patent documents from TRIAD and TETRAD, the main subdomains of interests can be outlined for the main countries/region, which together shared 93.5% of worldwide patenting. Semiconductors, Macromolecular Chemistry and Pharmacy-Cosmetics subdomains characterized the US technological development; the EU showed interest in Macromolecular Chemistry, Materials-Metallurgy and Pharmacy-Cosmetics; Japan stood out in Semiconductors, Electrical Components and Materials-Metallurgy; Korea paid attention to Electrical Components and Semiconductors; and China had an extremely small number of documents in the TRIAD and TETRAD. They could be highlighted in Technical Processes and Materials-Metallurgy. China became the main worldwide patentee in nanotechnology after 2007 and was the only country to increase the annual number of patents until 2010. China shared a small number of patent documents in USPTO, TRIAD and TETRAD and raised an issue whether their nanotechnologies are relevant economically or not. Other countries/region showed a recent decline in the number of patent documents per year, including the USPTO, TRIAD and TETRAD perspectives and this may be related to several factors, including the current economic recession, the delay in PCT processes and indexing, or even uncompleted data. The annual growth rate and the cumulative number of patents from 1995 to 2010 suggested that

nanotechnology development has achieved its initial stage of maturation. However, some biases were considered, such as data incompleteness, the need to forecast the nanotechnology subfield, and the link between nanoscience and nanotechnological development although considering the current paradigms. Moreover, changes in the developed forecasting curves or indicators cannot be predicted and monitoring the technological activities should be constantly reviewed in order to confirm evidence and test hypotheses that emerged.

Acknowledgements

The authors are grateful to the Brazilian National Council for Technological and Scientific Development (process number 160087/2011-2), the São Paulo Research Foundation (process number 2012/16573-7) and the Graduate Program in Materials Science and Engineering at the Federal University of São Carlos for supporting this work.

References

Alencar, M. S. M., Porter, a. L., & Antunes, a. M. S. (2007). Nanopatenting patterns in relation to product life cycle. *Technological Forecasting and Social Change*, *74*(9), 1661–1680. doi:10.1016/j.techfore.2007.04.002

Chen, H., Roco, M. C., Li, X., & Lin, Y. (2008). Trends in nanotechnology patents. *Nature nanotechnology*, *3*(3), 123–5. doi:10.1038/nnano.2008.51

Cheng, A., & Chen, C. (2008). The technology forecasting of new materials: the example of nanosized ceramic powders. *Romanian Journal of Economic Forecasting*, *4*, 88–110.

Dang, Y., Zhang, Y., Fan, L., Chen, H., & Roco, M. C. (2010). Trends in worldwide nanotechnology patent applications: 1991 to 2008. *Journal of nanoparticle research*, *12*(3), 687–706. doi:10.1007/s11051-009-9831-7

Galembeck, F., & Rippel, M. M. (2006). Nanotecnologia: estratégias institucionais e de empresas. In Strategic studies*: nanotechnology* (pp. 6–120). Brasília: Secretariat of Strategic Affairs. (Portuguese).

Glänzel, W., Debackere, K., & Meyer, M. (2007). "Triad" or "Tetrad"? On Global Changes in a Dynamic World. *SSRN Electronic Journal*. doi:10.2139/ssrn.1101439

Hu, A. G., & Jefferson, G. H. (2009). A great wall of patents: What is behind China's recent patent explosion? *Journal of Development Economics*, *90*(1), 57–68. doi:10.1016/j.jdeveco.2008.11.004

Huang, C., Notten, A., & Rasters, N. (2011). Nanoscience and technology publications and patents: a review of social science studies and search strategies. *The Journal of Technology Transfer*, *36*(2), 145–172. doi:10.1007/s10961-009-9149-8

Huang, Z., Chen, H., Chen, Z., & Roco, M. C. (2004). International nanotechnology development in 2003: Country, institution, and technology field analysis based on USPTO patent database. *Journal of Nanoparticle Research*, *6*(4), 325–354. doi:10.1007/s11051-004-4117-6

Huang, Z., Chen, H., Yip, A., Ng, G., Guo, F., Chen, Z., & Roco, M. C. (2003). Longitudinal patent analysis for nanoscale science and engineering : Country , institution and technology field. *Journal of Nanoparticle Research*, *5*, 333–363.

Hullmann, A., & Meyer, M. (2003). Publications and patents in nanotechnology An overview of previous studies and the state of the art. *Scientometrics*, *58*(3), 507–527.

Igami, M. (2008). Exploration of the evolution of nanotechnology via mapping of patent applications. *Scientometrics*, *77*(2), 289–308. doi:10.1007/s11192-007-1973-8

Kostoff, R. N., Koytcheff, R. G., & Lau, C. G. Y. (2007). Global nanotechnology research literature overview. *Technological Forecasting and Social Change*, *74*(9), 1733–1747. doi:10.1016/j.techfore.2007.04.004

Lv, P. H., Wang, G.-F., Wan, Y., Liu, J., Liu, Q., & Ma, F. (2011). Bibliometric trend analysis on global graphene research. *Scientometrics*, *88*(2), 399–419. doi:10.1007/s11192-011-0386-x

Martino, J. P. (Ed.). (1993). *Technological Forecasting for Decision Making*. New York: Mcgraw-Hill.

Milanez, D. H. (2011). *Nanotechnology: technological indicators on advances in materials based on patent analysis*. (Master's thesis) Federal University of Sao Carlos, Sao Carlos. (Portuguese). Retrieved January 28, 2013 from: http://200.136.241.56/htdocs/tedeSimplificado/tde_busca/arquivo.php?codArquivo=4653

Milanez, D. H., Amaral, R. M., Faria, L. I. L. & Gregolin, J. A. R. (2013). Assessing nanocellulose developments using science and technology indicators. *Materials Research.* Epub March 05, 2013. Retrieved April 19, 2013, from http://www.scielo.br/scielo.php?script=sci_arttext&pid=S1516-14392013005000033&lng=en&tlng=en.

Mogee, M. E. (1997). Patents and Technology Intelligence. In Ashton, W. B. & Klavans, R. A (Eds.) *Keeping abreast of science and technology: technical intelligence for business* (pp. 295-336). Columbus: Battelle Press.

Observatorynano. (2012). *Public Funding of Nanotechnologies* (pp. 1–22). Retrieved January 28, 2013 from: http://www.observatorynano.eu/project/filesystem/files/PublicFundingofNanotechnologies_March2012.pdf

OECD. (2009). *OECD Patent Statistics Manual*. Retrieved January 28, 2013from: http://dx.doi.org/ 10.1787/9789264056442-en

Observatoire des Sciences et des Techniques. (2010). *Science & technologie indicateurs*. Paris: Economica.

Porter, A. L., Youtie, J., Shapira, P., & Schoeneck, D. J. (2008). Refining search terms for nanotechnology. *Journal of Nanoparticle Research*, *10*(5), 715–728. doi:10.1007/s11051-007-9266-y

Roco, M. C. (2005). International Perspective on Government Nanotechnology Funding in 2005. *Journal of Nanoparticle Research*, *7*(6), 707–712. doi:10.1007/s11051-005-3141-5

Salerno, M., Landoni, P., & Verganti, R. (2008). Designing foresight studies for Nanoscience and Nanotechnology (NST) future developments. *Technological Forecasting and Social Change*, *75*(8), 1202–1223. doi:10.1016/j.techfore.2007.11.011

Scheu, M., Veefkind, V., Verbandt, Y., Galan, E. M., Absalom, R., & Förster, W. (2006). Mapping nanotechnology patents: The EPO approach. *World Patent Information*, *28*(3), 204–211. doi:10.1016/j.wpi.2006.03.005

United States. (2012). The National Nanotechnology Initiative. *NNI Supplement to the President's 2013 Budget*. Retrieved January 28, 2013 from: http://www.nano.gov/node/748

Wang, G., & Guan, J. (2012). Value chain of nanotechnology: a comparative study of some major players. *Journal of Nanoparticle Research*, *14*(2). doi:10.1007/s11051-011-0702-7

THE PATTERNS OF INDUSTRY-UNIVERSITY-GOVERNMENT COLLABORATION IN PHOTOVOLTAIC TECHNOLOGY

Huei-Ru Dong[1], Dar-Zen Chen[2] and Mu-Hsuan Huang[3*]

[1] *d99126002@ntu.edu.tw*
Department of Library and Information Science, National Taiwan University, Taiwan

[2] *dzchen@ntu.edu.tw*
Department of Mechanical Engineering and Institute of Industrial Engineering, National Taiwan University, Taiwan

[3] *Corresponding Author: mhhuang@ntu.edu.tw*
Department of Library and Information Science, National Taiwan University, Taiwan

Abstract
This research aims to understand the patterns of the Industry-University-Government (IUG) collaboration relationship. The degree of the involvement of the three sections in photovoltaic technology is observed from the co-authored patents on I-U-G collaboration retrieved from USPTO database between 2002 and 2011. This study hopes to determine the linkage between technical development and potential matches for institutional collaborations. The researcher first analyses and compares the number of co-authored patents and the share of patents contributed by each I-U-G sectors. Next, to understand how the institution, the university and the government participated in the development of photovoltaic technology, the study identifies the number of patents co-authored by two of the I-U-G sectors and the percentages of the co-authored patents over the total. Lastly, the main participants in U-I, U-G, and I-G collaborations are identified through close examination of co-authored patents on photovoltaic technology. The results reveal that industry has the highest shares both in the number of patents and of participating institutions; however, the percentage in co-authored patents of the industry is the lowest. On the other hand, the university has the highest shares of co-authored patents and of participating institutions.

Conference Topic
Technology and Innovation Including Patent Analysis (Topic 5) and Collaboration Studies and Network Analysis (Topic 6).

Introduction
Co-authorship signifies the collaborative relationship in scientific collaboration, including interaction between theoretical knowledge and technical data (Heffiner, 1981). Patent collaboration is as one of the key methods used to measure the

output of innovation system. Previous studies that investigated patent collaborations in national and regional innovation systems developed a strong dependency on foreign knowledge (Gao, Guan, & Rousseau, 2011; Chen & Guan, 2011). Ortega (2011) examines the collaborative patterns in the networks of patents and finds that the national collaboration can strongly and effectively transfer the patents.

Photovoltaic technology is a key research topic in the field of energy technology. As energy crisis intensified, governments worldwide have promoted policies of clean energy and photovoltaic technology. As a result, the increased academic and industrial sections have devoted to the development of photovoltaic-related technology as encouraged by the government.

Particularly, academic circles have proposed several types of innovation models. The triple helix innovation pattern reflects the relationship among academia, the industry and the government in one country (Etzkowitz & Leydesdorff 1995). Also, the triple helix indicates a transformation in the relationship among university, the industry and the government. Of the three types of organizations, the industry generally leads the development and relies heavily on R&D and patents to a higher extent, which explains why industry cooperates frequently with universities and research organizations. University supports innovative development by providing trained human resources, scientific research results, and theoretical knowledge to the industry, enhancing its role of innovation in knowledge-based societies (Meyer, Sinilainen, & Utecht 2003). The triple helix innovation system theory emphasizes the interdependent and independent relationship among the university, the industry and the government (Etzkowitz & Leydesdorff 2000; Etzkowitz 2003).

Sábato and Mackenzi (1982) noted that triangle model guide the development of government policies with its technical research and production. Etzkowitz & Leydesdorff (2000) conceptualizes the model as analytical difference from the national systems of innovation approach. There are three typical models of triple helix configurations which reflect different relations among I-U-G. Etzkowitz & Leydesdorff (2000) further classified triple helix innovation patterns and find that the national innovation system encompasses academia and industry as well as directs the relations among them.

The development of national innovation systems depends on various factors such as historical situation, policy guidance, economic development and natural resources. Nowadays, most countries and regions are encouraged to attain the form of triple helix model. Leydesdorff & Meyer (2003) pointed out that there are three functionally different sub-dynamics in the knowledge-based innovation system: economic exchanges in the market, geographical variations, and the organization of knowledge. Dolfsma & Leydesdorff (2008) discussed the knowledge based economy and found that medium-tech industry has a greater contribution than high-tech industry on knowledge creation.

Patent is an open and available information resources to measure the inventive activities and the collaboration status of the university, the industry and the

government. Lee (1996) considered patent as an important technology output which influence the attitude of academia toward the university–industry cooperation. Besides, patent data contains standardized information that is related to new ideas and technological developments (Pilkington, Dyerson, & Tissier, 2002; Frietsch & Grupp, 2006).

Meyer et al. (2003) explored the collaboration relationship between the industry and the university by combining patent analysis with an inventor survey. The university-industry relation has been changing owing to new technologies developed from academic research. Altlan (1987) pointed out that R&D collaboration is an important mode in university-industry relations. Manjarres-Henriquez, Gutierrez-Gracia and Vega-Jurado (2008) found that the university-industry relation showed positive effect on scientific productivity of the university. In the context of partnerships between industry and academia, Leroy and Doerig (2008) stated that the ownership of intellectual property must also be considered in addition to scientific activities.

Collaboration is an inherent aspect of the research activity, as information exchange reinforces the discussion and the production of new knowledge (Katz & Martin 1997). Collaboration pattern plays an important role in the national innovation system. Though the interconnections among the university, the industry and the government in innovation system have been recognized, to measure the innovative contribution and collaboration status of the triple helix is rather difficult, since relevant technology and R&D input and output information are usually immeasurable.

Thus, the objective of this research is to study the Industry-University-Government (I-U-G) relationship via examining scientific collaboration patterns in photovoltaic technology. The photovoltaic technology is collected from USPTO database between 2002 and 2011. This research first analysed the number of co-authored patents in I-U-G and the amount of co-authored patents between I-U-G and other technologies. This research then analysed the number of co-authored institutions in I-U-G and the amount of co-authored institutions between I-U-G and other technologies. At last, key institutions among U-I, U-G and I-G technology collaborations were analysed to understand institution participation in the development of photovoltaic technology.

Methodology

This study utilizes patentometric methods to explore the collaboration patterns in photovoltaic technology. Patentometrics use objective statistics to observe quantitative and qualitative performance of a research topic. Through analysis conducted based on the indicators, one can understand the structure of technological production capacity, as well as the trends in technological development, which establishes common frames of reference for further research.

Data collection

The source of patent information used in this study is based on the patents retried through the database of United States Patents and Trademark Office (USPTO) with the United States Patent Classification System (USPC). Since US patents are considered as an epitome of the global technological development, patents application in the United States is a strategic action for most of the inventors and authors worldwide to maintain competitive. Compare with International Patent Classification system, the United States Patent Classification System is updated more frequently and provides more detailed information on relevant patents, reflecting the advancement and innovation of technologies more accurately.

To obtain patents relevant to photovoltaic technology, searches through work reports published by OECD, FEEM (Fondazione Eni Enrico Mattei) and WIPO (World Intellectual Property Organization), and identifies patents through keyword search and queries through USPC classification numbers were conducted to assimilate relevant key words in the field. Keywords such as solar cell, photovoltaic, PV, and USPC classification number such as 136/258, 136/252, 136/262, 136/263 are employed as query in the USPTO patent database, each patent checked by the expert. There are a total of 6,840 utility patents on photovoltaic technology between 2002 and 2011.

Analysis on the scientific collaboration patterns in photovoltaic technology is conducted is grouped into three types: the industry, the university, and the government. Based on the triple helix innovation patterns following up on studies of Leydesdorff (2003) and Park, Hong & Leydesdorff (2005), who used the Science Citation Index for computing the mutual information in three dimensions. All of the photovoltaic patents data set and assignees are compiled and organized under attribution to the industry-university- government relations. The titles of assignees containing the abbreviations UNIV or COLL are labelled as the university. Then assignees are labelled as the industry if their titles contain any of the following identifiers: CORP, INC, LTD, SA or AG. The assignees are identified as the government if they are the public research institutions with abbreviations such as NATL, NACL, NAZL, GOVT, MINIST, ACAD, INST, NIH, HOSP, HOP, EUROPEAN, US, CNRS, CERN, INRA, and BUNDES in their title.

Table 1 Number and share of patents and institutions in photovoltaic technology in different sectors

	Industry	*University*	*Government*	*Total*
Patent N	6,232 (91.11%)	403 (5.89%)	310 (4.53%)	6,840
Institution N	1,057 (83.36%)	137 (10.80%)	74 (5.84%)	1,268

Table 1 tabulates the number of patents and institutions in the three sectors – the industry, the university, and the government. Among the total of 6,840 patents on photovoltaic technology, 6,232 patents (91.11%) are produced by industry, 403

patents (5.89%) by the university, and 310 patents (4.53%) by the government. Among the total of 1,268 institutions for photovoltaic technology, 1,057 (83.36%) are from the industry, 137 (10.80%) from the university, and 74 (5.84%) from the government.

Result

Number and share of I-U-G co-authored collaboration in photovoltaic technology

This study first calculates the patent numbers of the four I-U-G collaboration forms and the shares of patents produced by the three types of institution. Among the 6,840 photovoltaic-related patents, 103 patents are of I-U-G collaborative authorship (1.51%). As shown in Table 2, among 103 patents collected, 62 are from University-Industry (U-I) collaboration, 25 Industry-Government (I-G) collaboration, 14 University-Government (U-G) collaboration, and 2 I-U-G collaboration.

As listed in Table 2, the categorization of I-U-G collaboration patents further shows that among the 6,840 photovoltaic-related patents, 91.11% are produced by the industry, 5.89% and 4.53% are from the university and the government respectively. From the perspective of the industry, the share of I-U-G collaboration patents is relatively low (1.43%). To be specific, while the share of U-I collaboration patents of industry is the highest in this section, the collaboration share of patents by the industry is still low at 0.99%.

Table 2 Number and shares of patents for the four types of I-U-G collaborations in photovoltaic technology

Collaboration type	U-I	I-G	U-G	I-U-G	total
Patent N	62	25	14	2	103
Industry (91.11%= 6,232/6,840)	0.99%	0.40%	-	0.03%	1.43%
University (5.89%= 403/6,840)	15.38%	-	3.47%	0.50%	19.35%
Government (4.53%= 310/6,840)	-	8.06%	4.52%	0.65%	13.23%

Number and share of I-U-G co-authored collaboration in photovoltaic technology in different types of institutions

Table 3 shows that most of the institutions that obtained patents are from the industry, followed by the university and the government. Within the assignees of 6,840 patents related to photovoltaic technology, there are 131 institutions participating in I-U-G collaboration. And among these institutions 60 (45.80%) are from the industry, 47 the university (35.88%), and 24 (18.32%) the government.

Though the numbers of institutions from university and government participating in I-U-G collaboration are relatively low, one third of the patents produced by the

university and the government are co-authored. The percentage of the shares of institutions participating in I-U-G collaboration from the university and the government are 34.31% and 32.43% respectively. Among the 1,057 industrial institutions that produced photovoltaic patents, only 60 institutions produced I-U-G co-authored patents, accounting for 5.68% of the overall institutions.

In the three types of I-U-G co-authorships, with 69 institutions, U-I collaboration ranks the highest, including 35 from the industry and 34 from the university. Forty-two institutions participated in I-G collaboration, including 25 from the industry and 17 from the government. Twenty-two institutions participated in U-G collaboration, including 12 from the university and 10 from the government. Six institutions participated in I-U-G collaborations, including 2 from the industry, the university and the government for each.

A further analysis on the types of I-U-G collaboration for different types of institutions shows that for the industry, the percentage of I-U-G collaboration has a low rate at 5.68%. The percentage of collaboration with the university is slightly higher at 3.31%, but the percentage of collaboration with government is even lower (2.37%). More than one third of the institutions participated in I-U-G collaboration (34.31%), including the main collaboration with the industry (24.82%), followed by the collaboration with the government (8.76%) Nearly one third of the institutions from the governments have participated in I-U-G collaboration (32.43%). These institutions of government have worked with the industry (22.97%), followed by collaboration with university (13.51%).

Table 3 Numbers and shares of patents in photovoltaic technology by three types of institutions in different types of I-U-G collaboration

Collaboration type (Patent N)	U-I (69)		I-G (42)		U-G (22)		I-U-G (6)			Total (131)		
Institution N	U	I	I	G	U	G	I	U	G	I	U	G
	34	35	25	17	12	10	2	2	2	60	47	24
Industry (83.36%= 1,057/1,268)	-	3.31%	2.37%	-	-	-	0.19%			5.68%		
University (10.80%= 137/1,268)	24.82%	-	-	-	8.76%	-		1.46%			34.31%	
Government (5.84%= 74/1,268)	-	-	-	22.97%	-	13.51%			2.70%			32.43%

The number of institutions participating in I-U-G collaboration in photovoltaic technology by year

This study tracks the trends of institutions participating in I-U-G collaboration annually. Figure 1 lists the numbers of the three types of photovoltaic institutions in I-U-G collaboration between 2002 and 2011. The number of institutions involved in I-U-G collaboration reached its highest in 2011, showing a growing trend toward I-U-G collaboration for institutions. In regard to the institutions in different sections, the number of institutions from the industry maintained the

highest among the three, except during 2009-2010. The trend line shows that the number of institutions from the university exceeds that of institutions from industry in 2010.
Comparison of the trend lines for the three types of institutions shows that the growth rate for the number of university institutions is the highest (R2=0.7367), the growth rate for the number of industrial institutions is approaching linear (R2=0. 5026), and the growth rate for the number of governmental institutions is relatively low (R2=0.4041).

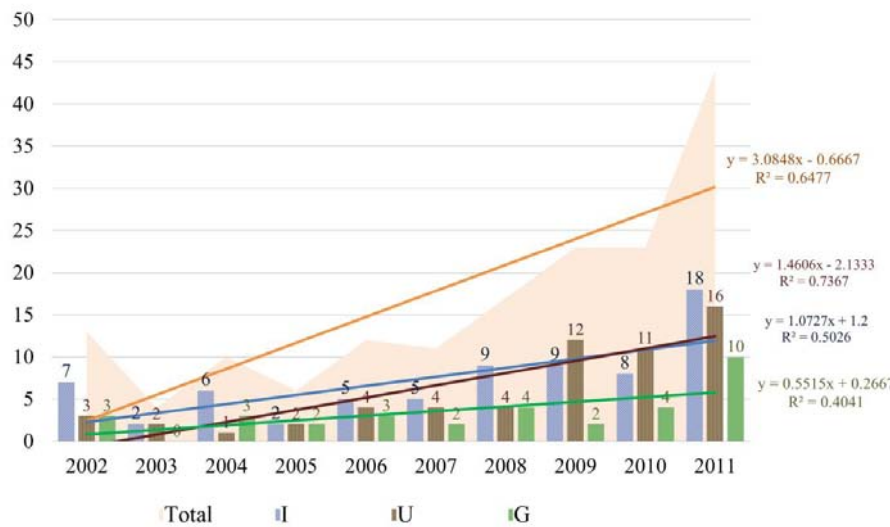


Figure 1 Number of institutions in I-G-U collaboration of photovoltaic patents by years

Key institutions in I-U-G collaboration

This study further conducts analysis to identify the key institutions participating in various types of I-U-G collaboration. The results are detailed as follows.

Key institutions in U-I collaboration

Table 4 lists the key institutions in U-I collaboration. Among these institutions, Tsing Hua University has produced the highest number of U-I co-authored patents. Among the 16 patents in photovoltaic technology, 15 belong to U-I collaboration (93.75%). Next, Hon Hai Precision Ind. Co., Ltd. obtained 14 patents from U-I collaboration (56.00%) out of its 25 patents in photovoltaic technology. The remained institutions have produced less than 10 patents from U-I collaboration.
Analysing the institutions in U-I collaboration from the perspective of countries, Japan obtained the highest number of institutions in U-I collaboration. Seven of the overall institutions are located in Japan, 6 in United States, 3 in South Korea,

2 in United Kingdom, and one in China, Taiwan, and France for each country respectively.

A closer observation on the types of institutions participating in U-I collaboration shows that 10 institutions are from the university and 11 are from the industry. From the perspective of patent number and the share in I-U-G collaboration in photovoltaic technology, universities have produced most patents from U-I collaboration including Tsing Hua University (15/16=93.75%), Kyoto University (2/3=66.67%), University Of Southern California (8/18=44.44%), Hanyang University, Osaka University, Seoul National University, and St. Andrews University (2/2=100%). Institutions from the industry include Hon Hai Precision Ind. Co., Ltd. (14/25=56%), Universal Display Corporation (8/20=40%), and Dow Corning Corporation and Isis Innovation Limited (2/2=100%).

Table 4 Key institutions in U-I collaboration

No.	Institution	Country	Type	I-U-G collaboration patent N	patent N	share
1	Tsing Hua University	China	U	15	16	93.75%
2	Hon Hai Precision Ind. Co., Ltd.	Taiwan	I	14	25	56.00%
3	The University Of Southern California	USA	U	8	18	44.44%
3	Universal Display Corporation	USA	I	8	20	40.00%
5	Samsung Electronics Co., Ltd.	South Korea	I	4	189	2.12%
6	California University	USA	U	3	29	10.34%
6	Princeton University	USA	U	3	52	5.77%
8	Dow Corning Corporation	USA	I	2	2	100.00%
8	Hanyang University	South Korea	U	2	2	100.00%
8	Isis Innovation Limited	UK	I	2	2	100.00%
8	Osaka University	Japan	U	2	2	100.00%
8	Seoul National University	South Korea	U	2	2	100.00%
8	St. Andrews University	UK	U	2	2	100.00%
8	Kyoto University	Japan	U	2	3	66.67%
8	Ecole Polytechnique	France	U	2	10	20.00%
8	Shin Etsu Chemical Co., Ltd.	Japan	I	2	11	18.18%
8	Rohm Company Limited	Japan	I	2	16	12.50%
8	Pioneer Corporation	Japan	I	2	17	11.76%
8	Idemitsu Kosan Co., Ltd.	Japan	I	2	40	5.00%
8	Hewlett-Packard Development Company, L.P.	USA	I	2	51	3.92%
8	Seiko Epson Corporation	Japan	I	2	86	2.33%

The university institution with high number of photovoltaic patents but low number of U-I collaboration patents is Princeton University (3/52=5.77%). Similar industrial institutions include Samsung Electronics Co., Ltd. (4/189=2.12%), Idemitsu Kosan Co., Ltd. (2/40=5%), Hewlett-Packard

Development Company, L.P. (2/51=3.92%), and Seiko Epson Corporation (2/86=2.33%). These figures show that university is more dependent on U-I collaboration than industry in U-I collaboration photovoltaic patent output.

Key institutions in I-G collaboration

Table 5 lists the key institutions involved in I-G collaboration. Among these institutions, Agency of Industrial Science & Technology has produced the most of the patents from I-G collaboration. Out of the 19 photovoltaic patents, 4 are I-G collaboration patents (21.05%). National Research Council of Canada has 3 I-G collaboration patents out of the 10 photovoltaic patents (30%). And the third is Xerox Corporation, with 3 I-G collaboration patents out of 62 photovoltaic patents (4.84%).

From the perspective of countries of which the 11 key institutions of I-G collaboration are located, 3 are located in South Korea, 2 in France and Belgium for each, and one in Japan, Canada, USA, and Germany respectively.

As for the types of institutions, 6 are from the government sector and 5 are from the industry. From the perspective of number and percentage of patents in I-U-G collaboration, in the industry, institutions produced higher number of patents from I-G collaboration. The institutions include Xerox Corporation (3/62=4.84%) and Samsung Electronics Co., Ltd. (2/189=1.06%).

Table 5 Key institutions in I-G collaboration

No.	Institution	Country	Type	I-U-G collaboration patent N	patent N	share
1	Agency of Industrial Science & Technology	Japan	G	4	19	21.05%
2	National Research Council of Canada	Canada	G	3	10	30.00%
2	Xerox Corporation	USA	I	3	62	4.84%
4	Framatome	France	I	2	2	100.00%
4	Office National Dapos;Etudes Et De Recherches Aerospatiales	France	G	2	2	100.00%
4	Umicore NV	Belgium	I	2	2	100.00%
4	Hanwha Chemical Corporation	South Korea	I	2	3	66.67%
4	Korea Institute of Science and Technology	South Korea	G	2	6	33.33%
4	Fraunhofer-Gesellschaft Zur Foerderung der Angewandten Forschung E.V.	Germany	G	2	13	15.38%
4	Imec Vzw	Belgium	G	2	13	15.38%
4	Samsung Electronics Co., Ltd.	South Korea	I	2	189	1.06%

Institutions from the industry that produced high number of photovoltaic patents but low number of patents from I-G collaboration include Hanwha Chemical Corporation (2/3=66.67%) and Framatome and Umicore NV (2/2=100%). As for institutions from the government, Agency of Industrial Science & Technology

(4/19=21.05%), National Research Council of Canada (3/10=30%), Office National Dapos; Etudes Et De Recherches Aerospatiales (2/2=100%), and Korea Institute of Science and Technology (2/6=33.33%) are included. The trend shows that, for I-G collaboration photovoltaic patent output, governmental institutions are more dependent on I-G collaboration than on industrial institutions.

Key institutions in U-G collaboration

Table 6 lists the key institutions in U-G collaboration. Centre National de La Recherche Scientifique - Cnrs and Imec Vzw are the institutions that have produced the highest number of patents from U-G collaboration. Each of the institutions has 3 patents. The remaining institutions have produced 2 for each. And from the perspective of countries where these institutions are located, 2 are from France and Belgium respectively, and one is from Singapore.

A closer look at the types of institutions in U-G collaboration shows that, of these institutions, 3 are from the government, and 2 are from the university. From the perspective of the number and percentage of patents in I-U-G collaboration, Centre National de La Recherche Scientifique - Cnrs (3/5=60%) and Agency for Science, Technology and Research (2/5=40%) have higher number of patents from U-G collaboration in the government sector. Universite Catholique de Louvain and Universite Peirre Et Marie Curie (2/2=100%) have the highest number of patents from U-G collaboration from the university sector.

Only one institution, Imec Vzw, produced high number of photovoltaic patents but low number of patents from U-G collaboration (3/13=23.08%). This shows that in comparison with the government, the output of photovoltaic patent is more dependent on U-G collaboration for the university.

Table 6 Key institutions in U-G collaboration

No.	Institution	Country	Type	I-U-G collaboration patent N	patent N	share
1	Centre National de La Recherche Scientifique - Cnrs	France	G	3	5	60.00%
1	Imec Vzw	Belgium	G	3	13	23.08%
3	Universite Catholique de Louvain	Belgium	U	2	2	100.00%
3	Universite Peirre Et Marie Curie	France	U	2	2	100.00%
3	Agency for Science, Technology and Research	Singapore	G	2	5	40.00%

Conclusion & Discussion

This research studies the scientific collaboration patterns in photovoltaic technology from USPTO database between 2002 and 2011, in order to examine the scientific collaboration pattern of the I-U-G relationship of photovoltaic technology. The authors analysed the numbers and the shares of co-authored patents between I-U-G, the number of co-authored institutions in I-U-G and the

amount of co-authored institutions between I-U-G. Key institutions among U-I, U-G and I-G technology collaborations are then identified to understand which institutions mainly participate in the development of photovoltaic technology.

The results show that the industry has produced the most number of patents, but the industry's involvement in I-U-G collaboration is the lowest (1.43%). On the other hand, though the university and the government have not produced as many patents in photovoltaic technology relatively, their involvements in I-U-G collaboration reach higher rates at 19.35% and 13.23% respectively. Moreover, the two have predominantly collaborated with industry. The trends show that I-U-G collaboration is an important route for cooperation on technological development for university and government. By contrast, I-U-G collaboration is clearly not as importance to the industry.

Majority of institutions that provide photovoltaic patents are from the industry, though their involvement in I-U-G collaboration is the lowest (5.87%). On the other hand, the number of institutions from university and government is relatively small, but these institutions have been further involved in I-U-G collaboration with at the rates of 34.31% and 32.43%. Again, these institutions also collaborated with the industry. A closer look at the number of different types of institutions of each year shows that the number of institutions from university has exceeded that of institutions from the industry since 2010.

From the institution's perspective, about one third of institutions from both university and government have been involved in I-U-G collaboration, showing that institutions from university and government are more receptive to I-U-G collaboration, while the willingness for I-U-G collaboration from the industry is low by comparison. The industry's frequent collaboration with the university or the government in the early stage of research and development may be part of the reasons. Also, the industry tends to develop independent technology. However, further studies are required to find out specific reasons for this phenomenon.

The key institutions in photovoltaic technology in I-U-G collaboration can be divided into two types. First, the institution focuses on I-U-G collaboration; the photovoltaic patents produced by these institutions are mostly the results of I-U-G collaboration. The second type is the institutions with high number of photovoltaic patents but low number of patents from I-U-G collaboration. These institutions may have participated in I-U-G collaboration, but they moved to the development for photovoltaic patents individually. Further analyses such as cited performance on the collaboration partners of these key institutions can help understand the patterns that encourage I-U-G collaboration, as well as help to gain more in-depth understanding to I-U-G collaboration.

References

Atlan, T. (1987). Bring Together Industry and University Engineering Schools, "In Getting More Out of R&D and Technology". The Conference Board, Research Report No. 904.

Chen, Z., & Guan, J. (2011). Mapping of biotechnology patents of China from 1995–2008. Scientometrics, 88(1), 73-89.
Dolfsma, W., & Leydesdorff, L. (2008). Medium-tech' industries may be of greater importance to a local economy than High-tech' firms: New methods for measuring the knowledge base of an economic system. Medical Hypotheses, 71(3), 330-334.
Etzkowitz, H. & Leydesdorff, L. (2000). The dynamics of innovation: from national systems and "Mode 2" to a Triple Helix of university-- industry-- government relations. Research Policy, 29, 109–123.
Etzkowitz, H. (2003). Innovation in Innovation: The Triple Helix of University-- Industry-- Government Relations. Social Science Information, 42(3), 293-337.
Etzkowitz, H., & Leydesdorff, L. (1995). The triple helix of university-- industry-- government relations: A laboratory for knowledge-based economic development. EASST Review, 14(1), 14-19.
Frietsch, R., & Grupp, H. (2006). There is a new man in town: the paradigm shift in optical technology. Technovation, 26(1), 463–472.
Gao, X., Guan, J., & Rousseau, R. (2011). Mapping collaborative knowledge production in China using patent co-inventorships. Scientometrics, 88(2), 343-362.
Heffner, A. (1981). Funded research, multiple authorship, and subauthorship collaboration in four disciplines. Scientometrics, 3(1), 5-12.
Katz, J. S., & Martin, B. R. (1997). What is research collaboration? Research Policy, 26(1), 1-18.
Lee, Y. S. (1996). Technology transfer and research university: A search for the boundaries of university industry collaboration. Research Policy, 25(6), 843-863.
Leroy, D., & Doerig, C. (2008). Drugging the Plasmodium kinome: the benefits of academia-industry synergy. Trends in Pharmacological Sciences, 29(5), 241-249.
Leydesdorff, L., & Meyer, M. (2003). The Triple Helix of university-industry-government relations. Scientometrics, 58(2), 191–203.
Manjarrés-Henríquez, L., Gutiérrez-Gracia, A., & Vega-Jurado, J. (2008). Coexistence of university-industry relations and academic research: Barrier to or incentive for scientific productivity. Scientometrics, 76(3), 561-576.
Meyer, M., Sinilainen, T., & Utecht, J. T. (2003). Towards hybrid Triple Helix indicators: A study of university-related patents and a survey of academic inventors. Scientometrics, 58(2), 321-350.
Ortega, J. L. (2011). Collaboration patterns in patent networks and their relationship with the transfer of technology: the case study of the CSIC patents. Scientometrics, 87(3), 657-666.
Park, H. W., Hong, H. D., & Leydesdorff, L. (2005). A comparison of the knowledge-based innovation systems in the economies of South Korea and the Netherlands using Triple Helix indicators. Scientometrics, 65(1), 3–27.

Pilkington, A., Dyerson, R., & Tissier, O. (2002). The electric vehicle: patent data as indicators of technological development. World Patent Information, 24(1), 5–12.
Sábato, J. & Mackenzi, M. (1982). La Producción de Tecnología. Autónoma o Transnacional. Maxico: Nueva Imagen.

PERFORMING INFORMETRIC ANALYSIS ON INFORMATION RETRIEVAL TEST COLLECTIONS: PRELIMINARY EXPERIMENTS IN THE PHYSICS DOMAIN (RIP)

Tamara Heck[1] and Philipp Schaer[2]

[1] *tamara.heck@uni-duesseldorf.de*
Heinrich-Heine-Universität Düsseldorf, Universitätsstr. 1, 40225 Düsseldorf (Germany)

[2] *philipp.schaer@gesis.org*
GESIS – Leibniz Institute for the Social Sciences, Unter Sachsenhausen 6-8, 50667 Köln (Germany)

Abstract
The combination of informetric analysis and information retrieval allows a twofold application. (1) While informetrics analysis is primarily used to gain insights into a scientific domain, it can be used to build recommendation or alternative ranking services. They are usually based on methods like co-occurrence or citation analyses. (2) Information retrieval and its decades-long tradition of rigorous evaluation using standard document corpora, predefined topics and relevance judgements can be used as a test bed for informetric analyses. We show a preliminary experiment on how both domains can be connected using the iSearch test collection, a standard information retrieval test collection derived from the open access arXiv.org preprint server. In this paper the aim is to draw a conclusion about the appropriateness of iSearch as a test bed for the evaluation of a retrieval or recommendation system that applies informetric methods to improve retrieval results for the user. Based on an interview study with physicists, bibliographic coupling and author-co-citation analysis, important authors for ten different research questions are identified. The results show that the analysed corpus includes these authors and their corresponding documents. This study is a first step towards a combination of retrieval evaluations and the evaluation of informetric analyses methods.

Conference Topic
Old and New Data Sources for Scientometric Studies: Coverage, Accuracy and Reliability (Topic 2), Research Fronts and Emerging Issues (Topic 4) and Open Access and Scientometrics (Topic 10).

Introduction
Informetric analyses are generally used to gain insights into a scientific domain and to better understand scholarly activities. A common approach is the use of statistical modelling or visualization techniques to get a more profound overview of a scientific domain or a specific topic. The process of science modelling tries to describe and formalize these approaches. Examples for methods that are used in

the science modelling community are co-occurrence or co-authorship analyses or bibliographic coupling (see Scharnhorst, Börner & Besselaar, 2012). While in most cases these models are used to make scientific rankings or to draw so-called science maps, some approaches try to combine science modelling and information retrieval (IR) research (Mutschke et al., 2011).

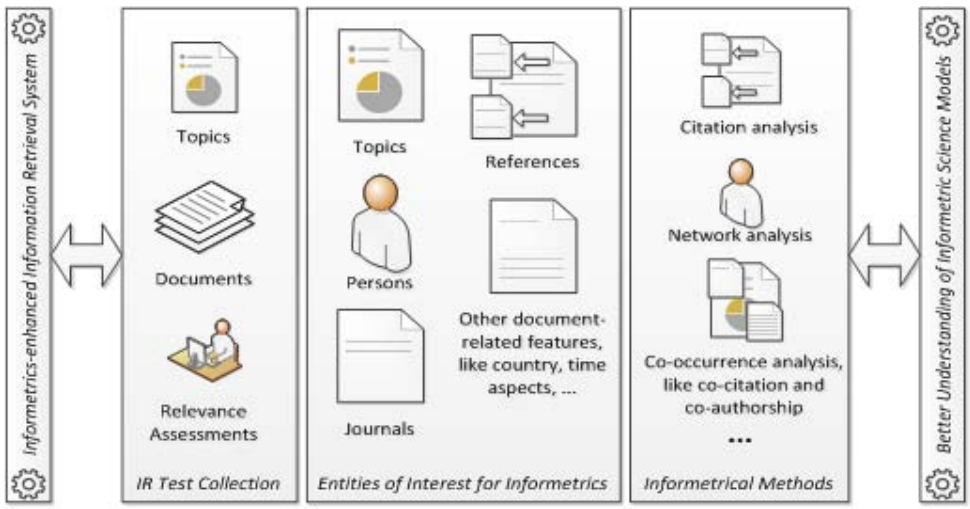


Figure 1. Mutual benefits of IR test collections and informetric analysis methods.

Authors like Ingwersen (2012) propose a more application-driven view on informetrics. The main idea is that a more profound insight into a science system can be exploited to support the search process in a scholarly information system. Entities that are usually observed are authors, topics or publication organs like journals, publishers etc. The more we know about the different entities and their connection to each other in the scientific publication system, the more we can use this information to enrich the retrieval process. A classic example of this is the Bradfordizing method proposed by White (1981), where typical power-law distributions in bibliographic data sets are used to offer a different ranking mechanism. Up to then Bradford's Law was only used to detect core journals in a scientific field but this qualitative information was not used in actual retrieval systems. It was clearly shown that highly co-occurring attributes have a strong selectivity and can be applied as a ranking weight which can lead to different view on the document space (Schaer, 2011).

While in the IR community a decades-long evaluation tradition exists (with evaluation campaigns like TREC[131] or CLEF[132]), informetrics lacks this kind of tradition. To overcome this gap Mutschke et al. (2011) proposed the use of standard IR evaluation methods as a test-bed and "litmus test" for science models.

[131] http://trec.nist.gov
[132] http://www.clef-initiative.eu

The interconnection between IR test collections, IR evaluation methodologies and informetrics is shown in figure 1. The overall idea is an informetrics-enhanced IR system that incorporates all the previous elements to complement existing approaches through a deeper understanding of informetric science models.
The following paper describes the outcomes of a preliminary experiment on analysing a standard IR evaluation collection. We used the iSearch[133] test collection as well as data derived from an earlier experiment and an interview series with physicists. The information from the experiment is used to cross-check whether (a) the specific topics the physicists were interested in and (b) the important authors identified during the interviews are included in the iSearch corpus. If the results are positive, this standard IR corpus might serve as a basis for further retrieval and science model testing.
In the next section we will describe both data sets we want to map: the iSearch test collection and the topics and important authors extracted from social information and informetric analyses as well as interviews. Thereafter the results of the mappings are summarised and we will discuss the outcomes of this preliminary experiment in the final section.

Data Sets

The iSearch Test Collection

The iSearch test collection (Lykke et al., 2010) consists of the three standard parts of an IR test collection: (1) a corpus of documents, (2) a set of topics, and (3) relevance assessments. The corpus consists of documents from the physics domain: 18,222 monographic library records, 160,168 scientific papers and journal articles in PDF full texts and their corresponding metadata, as well as 274,749 abstracts with their corresponding metadata. Additionally the data set includes more than 3.7 million extracted internal citations. The monographic records were extracted from the Danish National Library and the full text and metadata sets were crawled from the arXiv.org open-access/preprint repository. The set of 65 topics and their relevance assessments (~200 per topic) were extracted from 23 lecturers, PhD and MSc students from three physics departments. Up to now this test collection was mainly used in the domain of contextual and task-based IR research because of the rich and realistic search tasks that allow an in-depth analysis of user intentions and expectations in the retrieval task.
The deposition of article preprints at arXiv.org is usually best practice in most physics working groups, and so we see potential in the document corpus itself because of the size and coverage. This corpus is a rich document set from a scientific domain (in contrast to other data sets without the scientific and discipline relatedness, like the typical TREC data sets) and includes everything needed to carry out an IR evaluation (in contrast to other scientific literature

[133] http://itlab.dbit.dk/~isearch/?q=node/1

corpora like INSPEC, Scopus or the Web of Science). On the other hand an author who deposits his paper is only supposed to provide a minimal set of (unstructured) metadata. In fact there are very few instructions and rules on how to enter the metadata. This results in a large but very heterogeneous document set.

Topics and Important Authors in the Physics Domain

To find out which resources are relevant for a user and which are not, user feedback and relevance judgements are needed. iSearch only includes such judgements for documents, but since we want to focus on important authors we need additional information. In our approach we obtained this user information from semi-structured interviews. These evaluations were part of a project to recommend collaboration partners to ten participating scientists (Heck, Peters & Stock, 2011). The aim was to recommend to a researcher authors who have similar research interests and thus could be potential collaborators. Therefore the interviewees should state whether the recommended authors are important for their current research.

The author names the physicists should evaluate were extracted using social information data. In the Web of Science[134] author names were gained using bibliographic coupling of authors, i.e. authors who have many references in common with the target researchers (the physicists) are supposed to be similar. Thus their written papers might be relevant and they might be potential collaboration partners. In Scopus[135] those authors who were co-cited many times with our target researchers were extracted (White & Griffith, 1981). In CiteULike[136] those authors were supposed to be similar whose articles have tags (assigned by the service's users) in common with the target researchers' articles, or whose articles were bookmarked by users, who also bookmarked the target scientists' papers (collaborative filtering, see e.g. Marinho et al., 2011). These author names were rated by the target scientists on a scale from 1 (not relevant for current research) to 10 (highly relevant for current research). Furthermore each physicist described his research interests with specific terms during the interview.

Data Mapping in the iSearch Corpus

Design of the Experiment

To prove the assumptions formulated in the introduction we first have to test whether the iSearch corpus is an appropriate tool to do such experiments. One criterion is that the set includes articles written by the important authors identified in the interviews. If the physicist searches for literature in his research domain, he would expect to find articles that are very relevant for his research topic. Thus the articles should be written by those authors the physicist has claimed important for

[134] http:/www.webofknowledge.com
[135] http:/www.scopus.com
[136] http://www.citeulike.org/

his research. For the analysis of the iSearch corpus we use those authors the target physicists claimed important for their current research. We call them important authors. Important authors are those authors who in the evaluation process were rated with 5 or higher and who were explicitly named by the physicists. If the iSearch corpus includes articles from these authors – derived from methods like author-co-citation and bibliographic coupling – a retrieval system including information from science models could be evaluated on the basis of this corpus.

Table 1. Descriptions of research interests and research topics of the 10 physicists

Topic ID	*Description of research interest*
sci001	Modelling blood flow processes relating to viscosity and the formation of diseases. Analysing properties of polymers and microswimmers for medical obligations.
sci002	Biomolecular multiscale simulations concerning Alzheimer disease. Analysing protein aggregation and protein-protein interaction like amyloid ß-peptide.
sci003	Multiscale protein modelling and computational simulation. Analysing the properties and dynamics of fluids and polymers.
sci004	Interested in polymer catalysis and neutron scattering.
sci005	Analysing polymer-membrane interactions and the diffusion of red blood cells.
sci006	Spintronics in carbon nanostructures, carbon nanotubes and the raman spectroscopy.
sci007	Interested in photoelectron spectroscopy, (ferro) magnetic and electronic properties.
sci008	Simulation of crumpled elastic sheets and its mechanical deformation. Buckling of capsid proteins.
sci009	X-ray and neutron scattering in high-correlated electron systems and the building of instruments.
sci010	Analysing dynamics of glass-forming liquids. Interested in inelastic neutron scattering, dielectric spectroscopy and rheology. Doing simulations of polymers and other amorphous material.

In the interview each physicist described his research interests and research focuses with appropriate terms (see table 1). Our assumption is: If the physicist searches for literature he would probably use those terms as search terms he used to describe his research interest and research focus with. Thus the terms derived from the descriptions of the physicists' research focus (further described as topics sci001 – sci010) are used as search terms in the iSearch corpus. As some single terms would describe a research focus in a very common way, these terms are only used in combination; therefore the actual query composition was done manually based on the outcomes of the interviews to best reflect the physicists' interests. We indexed all available metadata (~453,000) and the full text data sets (~160,000) in the Solr[137] search engine and applied a standard Porter stemmer and

[137] http://lucene.apache.org/solr

an English stop-word list. The search is done using Solr's standard retrieval method, which is based on an extended Boolean model that allows the extraction of co-occurring entities (facets). We extracted all author names and the number of their articles from the retrieved documents. We only focused on authors and will leave the analysis of journals and references for future work.

Results

The physicists named 18 to 55 authors who are relevant for their current research (column 2 in table 2). We searched for those authors in the iSearch corpus. To secure correctness of the important authors and obviate author ambiguity, the author names were verified manually on the basis of co-authorship, article title and journal title. In our results we analysed two aspects, namely the important authors and their articles. We used three different data sets:

1. The whole iSearch corpus.
2. One subset per physicist, which was retrieved using the physicist's topic-describing terms by searching in title, abstract and full text (see previous section).
3. The top 50 documents of set number 2 ranked by Solr's TF*IDF implementation.

The left part of table 2 shows the coverage of the authors in the iSearch corpus, meaning at least one document of an important author can be found in the corpus. 199 of 287 unique important authors (IA) are in the corpus; on average they have nearly 19 articles in iSearch. For each topic at least 57% of IA are in the iSearch corpus. Sci006, with 29 named IA, even has 100% coverage. When using the terms describing the physicists' research interests as query terms (the subsets), nearly 70% of the previously named IA were included. But under the top 50 ranked articles there are on average only 4.8 IA. Of course the coverage depends on the time the corpus was created. Some interviewed physicists are rather novice researchers, i.e. they also named novice physicists, who are not in iSearch, as important for their research.

In the right part of table 2 we report on the coverage of the documents authored by the important authors (IAD) within the three described document pools. Note that the numbers of IAD in iSearch are approximated as it cannot be proved that every single document is really written by the correct important author. That means author ambiguity can be eliminated in the top 50 documents and for the most part in the subsets, but not in the corpus. Column 6 in table 2 (total docs in topic subset) shows the number of documents that are found with the topics showed in table 1. In these subsets the number of articles written by IA (column IAD in topic subset) ranges from just 11 documents for topic sci002 up to 426 for topic sci006 (avg. of 164.7). Within the top 50 documents on average only 4.8 IAD were included. For two topics (sci002 and sci007) no single IA or IAD were included in the top 50 documents.

Concerning both IA and IAD the coverage under the top 50 articles is weaker than in the total iSearch corpus and in the subsets. For example: Sci007 named 18

important authors. 16 of them are in the iSearch corpus. But a search with the query terms derived from the descriptions of the researcher's interests ranks no articles of IA under the top 50. Nevertheless in sci007 207 articles (IAD in topic subset) of 15 important authors (IA in topic subset) are found with the query terms determined by the physicist's research interest descriptions. Moreover, no correlation could be detected between the size of the topic subsets and the number of IA and IAD found within these subsets, i.e. you cannot state that the bigger the subset is, the more IA and IAD are included. E.g. sci003 has over 90,000 documents in the subset. All IA found in the total iSearch corpus are included in this subset, and about 86% of the IAD. But sci004, with only 16,042 documents, has similar results. Here all but one IA are included in the subset as well as 62% of the IAD. Sci008, which has more than twice as many documents as sci004, covers only about 65% of IA and about 20% of IAD. To summarize, it can be said that the coverage of IA and IAD is quite high in the iSearch corpus (nearly 70% of IA could be found) but in the subsets and especially in the top 50 ranked documents the coverage is quite low in most cases.

Table 2. Overview on the coverage of important authors (IA) and documents of important authors (IAD) within three different document pools: (1) the whole iSearch corpus, (2) a topical subset and (3) the top 50 TF*IDF ranked documents.

	Important authors				Documents of important authors			
Topic	y R e n i S e p i f n t c				i n t c n i S o p i i n t			
sci001	35	20	7	3	3152	291	24	3
sci002	27	17	7	0	1700	147	11	0
sci003	20	12	12	3	94205	142	123	2
sci004	24	17	16	5	16042	214	134	4
sci005	45	28	24	7	25169	299	185	12
sci006	29	29	28	13	61132	928	426	10
sci007	18	16	15	0	80846	283	207	0
sci008	55	34	22	10	39570	590	116	11
sci009	21	20	18	2	34814	723	274	1
sci010	21	14	14	3	57368	223	147	5
avg.	29.5	20.7	16.3	4.8	41399.8	384	164.7	4.8

Discussion and Future Work

We presented the outcomes of a preliminary experiment of mapping 10 specific scientific research interests onto the iSearch corpus. The statements of the physicists about authors being relevant for their current research are used as qualitative criterion to draw conclusions about the appropriateness of iSearch for evaluating informetric analyses.

Concerning the quite high coverage of important authors (nearly 70%), we assume that the iSearch corpus is appropriate for an evaluation of a retrieval approach that uses informetric methods to improve the retrieval process. In the future project we would like to use the references of the iSearch corpus and build

a retrieval system that also applies bibliographic coupling and co-citation analyses to improve the results for the user. For the evaluation we would use not only the external feedback by the physicists, but also the relevance feedback of the iSearch corpus. However the physicists' feedback are beneficial because they include concrete relevance ratings of important authors, which can be used to make further statements about the retrieval results. The relevance feedback in the iSearch corpus allow statements about the relevance of articles, but not about concrete authors. Both articles and authors might be important for a user who searches for relevant research literature in a retrieval system.

Concerning the coverage of important authors in the top 50 subset, we suppose that good retrieval results should rank the articles of the important authors at very high positions. The TD*IDF ranking alone doesn't seem to be powerful enough. It is assumed that methods like co-citation analysis and bibliographic coupling will improve both document and also author retrieval. It should be tested at which stages and in which processes of a retrieval system these approaches could be applied. One idea is to re-rank the documents, which were retrieved by e.g. co-word analysis. Depending on the users' need, informetric methods may also be applied before co-word approaches and ranking. The analysis of important journals is another method to gain more relevant articles.

References

Heck, T., Peters, I. & Stock, W.G.(2011). Testing collaborative filtering against co-citation analysis and bibliographic coupling for academic author recommendation. *Proceedings of the 3rd ACM RecSys'11 Workshop on Recommender Systems and the Social Web* (pp. 16–23). Chicago: ACM.

Ingwersen, P. (2012). Citations and references as keys to relevance ranking in interactive IR. In *Proceedings of the 4th Information Interaction in Context Symposium* (p. 1). New York: ACM.

Lykke, M., Larsen, B., Lund, H. & Ingwersen, P. (2010). Developing a test collection for the evaluation of integrated search. In C. Gurrin, Y. He, G. Kazai, U. Kruschwitz, S. Little, T. Roelleke, S. Rüger & K. Rijsbergen (Eds.) *Advances in Information Retrieval* (pp. 627–630). Berlin/Heidelberg: Springer.

Marinho, L.B., Nanopoulos, A., Schmidt-Thieme, L., Jäschke, R., Hotho, A., Stumme, G. & Symeonidis, P. (2011): Social tagging recommender systems. In F. Ricci, L. Rokach, B. Shapira & P.B. Kantor (Eds.) *Recommender Systems Handbook* (pp. 615–644). Berlin: Springer.

Mutschke, P., Mayr, P., Schaer, P. & Sure, Y. (2011). Science models as value-added services for scholarly information systems. *Scientometrics*, 89, 1, 349–364.

Schaer, P. (2011): Using lotkaian informetrics for ranking in digital libraries. In C. Hoare & A. O'Riordan (Eds.) *Proceedings of the ASIS&T European Workshop 2011.* Cork: ASIS&T.

Scharnhorst, A., Börner, K. & Besselaar, P. van den (Eds) (2012). *Models of Science Dynamics Encounters Between Complexity Theory and Information Sciences*. Berlin: Springer.
White, H.D. (1981). "Bradfordizing" search output: how it would help online users. *Online Information Review,* 5, 1, 47–54.
White, H.D. & Griffith, B.C. (1981). Author cocitation: A literature measure of intellectual structure. *Journal of the American Society for Information Science*, 32, 3, 163–171.

POSSIBILITIES OF FUNDING ACKNOWLEDGEMENT ANALYSIS FOR THE BIBLIOMETRIC STUDY OF RESEARCH FUNDING ORGANIZATIONS: CASE STUDY OF THE *AUSTRIAN SCIENCE FUND (FWF)*

Rodrigo Costas[1] and Alfredo Yegros-Yegros[2]

[1] *rcostas@cwts.leidenuniv.nl;* [2] *a.yegros@cwts.leidenuniv.nl*
CWTS-Centre for Science and Technology Studies, Leiden University, PO Box 905
2300 AX Leiden (the Netherlands)

Abstract
This paper presents a case study on the presence of funding acknowledgements among Austrian publications with a special focus on *Austrian Science Fund (FWF)*. Scientific publications funded by the FWF have been studied by means of the bibliographic records maintained by the funding organization and also by the presence of Funding Acknowledgements in the publications (FA analysis). It is observed that more than 50% of all publications funded by the FWF are detected only by means of the FA analysis, thus reinforcing the role of this type of analysis for bibliometric studies of funding organizations. Disciplinary differences have also been found, with the Social and Economic sciences showing the lowest rate of funding, but also the lowest rate of properly acknowledging their funding sources.

Conference Topic
Old and New Data Sources for Scientometric Studies: Coverage, Accuracy and Reliability (Topic 2) and Science Policy and Research Evaluation: Quantitative and Qualitative Approaches (Topic 3).

Introduction
Research funding organizations play an important role in scientific development and they have a strong interesting in studying their role and influence in the scientific landscape, particularly through bibliometric indicators. However, one of the challenges for bibliometric studies of funding organizations is how to collect reliable data on publications that reasonably can be linked to funded projects by these organizations, or as Hornbostel (2012) already stated "where data about funding should be collected: from the recipient or at the funding institution?".
In this paper we focus on the combination of two approaches that can help to solve this question, namely: the records maintained by the research funding institution and the publications where the authors (the recipients) acknowledged their funding sources. Acknowledgments and particularly funding acknowledgements (FA) are a common element in science (Tiew & Sen, 2002), as

authors of publications indicate through them the sources of funding or economic support with the research and the publication was made possible. On the other hand, funding organizations frequently demand researchers to inform them about the outputs of their funded projects, particularly publications, and may also keep records of papers that are assumed to result from their funding (Rigby, 2013).
In this paper we focus on the *Fonds zur Förderung der wissenschaftlichen Forschung* (FWF), the Austrian Science Fund, with the aim of determining how the authors have identified their funding by this organization, both through the bibliographic records maintained by the FWF and also by the FAs given in Austrian publications.

Objectives

The main objective of this paper is to study the differences and possibilities of two methods of data collection for the study of research funding organizations: the use of records detected/maintained by the funding organization and the analysis of the FA of scientific publications.

Methodology - our case study

This is a case study on the *Fonds zur Förderung der wissenschaftlichen Forschung* (FWF), the Austrian Science Fund, the central funding organization for basic research in Austria. The Austrian Science Fund (FWF) is Austria's central funding organization for basic research (http://www.fwf.ac.at/en/index.asp).

Publication data collection of the FWF

A bibliometric study has been recently developed at CWTS for the FWF (van Wijk & Costas-Comesaña, 2012). In that study, FWF provided CWTS with a list of publications from their own records (period 2001-2010). The publications supplied by the FWF were matched against the CWTS in-hose Web of Science database, based on bibliographic elements such as author names, publication year, journal title, etc. matching ~80% of all the publications initially supplied.
The van Wijk & Costas-Comesaña (2012) study was focused only on the Sciences and Social Sciences fields, excluding the Humanities (category 6 as defined by the OECD[138]). In this study we build upon that database and consider the same field delineation. The time period covers 2009-2010. For this study a manual check of the unmatched publications was performed in order to increase the quality of this data matching process. In table 1 the main figures related with the processing of the FWF records and their matching with the Web of Science are presented.

[138] http://www.oecd.org/science/innovationinsciencetechnologyandindustry/38235147.pdf. For the matching of Web of Science with the OECD categories we used an internal classification of journals to OECD fields available at CWTS.

Table 1. Main figures regarding the input data from the FWF records.

FWF System	*Records*
Total records with publication year 2009-2010 from FWF	**3198**
Matched in the WoS database	2806
Non-matched	392
Unique matched records	*Records*
Unique records matched in the WoS database	2580
Unique records matched in the WoS database (2009-2010)	**2437**

Table 1 shows that the initial input from the FWF system with publication year between 2009-2010 was composed by 3198 records. Around 88% of these records were effectively matched against the Web of Science. Considering the records matched, duplicated were removed and also some publications with a mismatch between the publication year in the FWF database and the Web of Science. The final set of publications was composed by 2437 unique WoS-covered publications.

Detection of variants of the FWF in the Funding Acknowledgements of publications

A second step was the identification of all the variants of the FWF mentions in the funding acknowledgments of scientific publications. This particular difficulty due to the low standardization of the funding bodies names (Rigby, 2011) and for this study this task has been performed manually. As a result more than 400 variants of the FWF were identified in the database.

Indicators

For some of the analysis, particularly citation analysis we have employed the CWTS standard methodology and indicators (Waltman et al, 2011a, 2011b). In this sense, the results presented in the citation analysis can be slightly different from that of the publication analysis (see Results Table 2 and following). The main indicators included in the analysis are: *P* (number of articles, letters and reviews; in the citation analysis letters have been weighted by 0.25), *TCS* (total number of citations received up to 2011, self-citations excluded), *MCS* (mean citation score), *MNCS* (mean normalized citation score, this is a measure of the impact of publications compared to the world citation average in the WoS subject categories of the publications), *MNJS* (mean normalized journal score, this is the impact of the journals in which publications are published, compared to the world citation average in the fields covered by these journals).

Results

Analysis of publications

In this section we present the main results regarding the analysis of the Austrian and FWF publications. Austria presents a total of 30362 publications (considering

all document types) during the period 2009-2010. However, when limited to the same fields that were considered in the FWF study (van Wijk & Costas, 2012 – excluding the Humanities fields) the final set of publications for the country is 29883.

In Table 2 we present the analysis of the presence of FA across different groups of publications, particularly taking into account the role of FWF in them.

Table 2. Coverage of FWF funded publications within Austrian publications in WoS period 2009-2010.

Matching	*Records*	*%*	*Observations*
Austrian WoS records	29883	100	
FWF records	2347	8	90 records didn't have an Austrian address
Non-FWF	27536	92	
Austrian papers with any FA	12202	41	
Austrian papers with FA to FWF	4146	34	% based on the Austrian papers with FA
FWF records	2347		
** With FA to FWF	1700	72	% of publications in the FWF system that include FA to the FWF
** Without FA to FWF	647	28	% of publications in the FWF system without FA to the FWF
Publications with FA to FWF but not in FWF records	2446		
Total FWF funded publications (FWF records without FA + Publications with FA to FWF but not in FWF records)	4793	100	
% in FWF records	2347	49	
% not in FWF records	2446	51	

Table 2 shows how FWF has funded ~8% of Austrian publications. Regarding the publications included in the FWF records 72% of them have in fact an FA to the FWF, while 28% of the publications covered in the FWF records do not hold such acknowledgement to the funder (although of course lately acknowledged by the inclusion of the publication in the FWF database). It is remarkable that 2446 publications have acknowledgements to the FWF but they are missing from the FWF records. Thus, if we consider all the publications that by any of the two approaches (i.e. FWF records or FA) show a funding relationship with the FWF we end up with 4793 publications. This means that ~51% of the funded publications by the FWF are missing from FWF records, while ~14% are detected only through their records but not through the FA analysis. This is also an interesting result, the fact that 14% of FWF publications (from the combination of FA and FWF records) do not have an acknowledgement suggest that somehow

the authors 'forget' at the time of publication that they should acknowledge the FWF. This would be then the lower bound of what we could consider the "FA forgetfulness", although in this account we would still be missing those publications that should have acknowledged some funding by the FWF but they didn't do it in any way (nor by FA neither reporting the FWF).

Citation analysis

In table 4, we compare the impact of the Austrian publications depending on the fact if they have funding or not.

Table 4. Funded vs. non-funded publications.

Publications	*P*	*% pubs*	*TCS*	*% cits*	*MCS*	*MNCS*	*MNJS*
Non-funded	10673.5	46.7	31416.69	34.1	2.94	0.97	0.92
Funded	12193.25	53.3	60816.96	65.9	4.99	1.51	1.35

In this table the number of funded publications outperforms that of the non-funded publications. It must be taken into account that some document types have been excluded for the citation analysis and that letters are weighted by 0.25. This means, as compared to table 2, that a substantial amount of publications without funding are document types not included in the citation analysis. Regarding the impact of the publications, publications with FA present a higher impact in all indicators, which is in line with previous studies (Costas & van Leeuwen, 2012).
In table 5 all publications that have detected to be funded by the FWF are compared to the rest of the country.

Table 5. FWF vs. non FWF publications in Austria

Publications	*P*	*% pubs*	*TCS*	*% cits*	*MCS*	*MNCS*	*MNJS*
Non FWF	18125.75	79	71206	77	3.93	1.23	1.10
Total FWF	4737	21	21009.75	23	4.44	1.37	1.32

Publications with funding from the FWF present in general a higher impact as compared to that of the rest of Austria. This is in line with the results by van Wijk & Costas-Comesaña (2012).
The next question is about the impact of the publications that show an FA to the FWF but are not recorded in the system of the FWF compared to those that are in the FWF records (table 6).

Table 6. FWF records vs. Only FA to FWF

Publications	*P*	*% pubs*	*TCS*	*%cits*	*MCS*	*MNCS*	*MNJS*
Only FA to FWF	2443	51.6	10352.0	49.3	4.24	1.39	1.32
FWF records	2294	48.4	10657.8	50.7	4.65	1.36	1.32

The publications that are retrieved only through their FAs amount to some more than 51% of all publications funded by the FWF. Again, differences with table 2 are explained by the limitation of document types that are considered for citation analyses (i.e. articles, reviews and weighted letters). Regarding the number and share of citations to both groups, the publications that are only retrieved through FAs attract around 49% of all the citations of the organization. The MNCS indicator shows that these publications are slightly more cited than those publications that are in the system of the FWF (when normalizing by publication year, document type and fields), although the differences are small.

Disciplinary analysis

In this section we present a distribution of publications across fields. We focus again on all document types with the intention of showing the presence of publication with and without funding (also with and without FWF participation) across fields. For the disciplinary classification we have used the Dutch NOWT (Dutch Observatory of Science and Technology - http://nowt.merit.unu.edu/) classification (Table 7).

Table 7. FA analysis by fields (Austria and FWF)

NOWT Discipline	*P*	*P with funding (any)*	*Pubs in FWF system*	*Pubs with FA to FWF*	*Pubs only FA to FWF*	*Pubs only in FWF system (1)*	*Total FWF*	*%gene-ral FA*	*%pubs only FA to FWF*	*%FA 'forget fulness'*
MEDICAL SCIENCES	14203	3686	424	757	460	127	884	26.0	52.0	14.4
CHEMISTRY, PHYSICS AND ASTRONOMY	6098	3857	816	1583	981	214	1797	63.3	54.6	11.9
LIFE SCIENCES	5734	3261	754	1206	599	147	1353	56.9	44.3	10.9
EARTH AND ENVIRONMENTAL SCIENCES	2451	1414	221	422	251	50	472	57.7	53.2	10.6
MATHEMATICS, STATISTICS AND COMPUTER SCIENCE	1830	948	275	518	349	106	624	51.8	55.9	17.0
ENGINEERING SCIENCES	1826	757	96	171	121	46	217	41.5	55.8	21.2
SOCIAL SCIENCES	1067	91	32	24	17	25	49	8.5	34.7	51.0
ECONOMICS, MANAGEMENT AND PLANNING	708	31	34	13	10	31	44	4.4	22.7	70.5
HEALTH SCIENCES	505	115	19	16	7	10	26	22.8	26.9	38.5
MULTIDISCIPLINARY JOURNALS	237	165	40	66	37	11	77	69.6	48.1	14.3

(1) Without FA to FWF but in FWF records.

Table 7 shows that the fields with more funding acknowledgements are Chemistry, Physics and Astronomy, Earth and Environmental Sciences, Life Sciences and Mathematics, Statistics and Computer Science, all of them with more than 50% of their publications acknowledging some kind of funding. On the other hand, the fields with the lowest levels of funding are those of the Social Sciences and Economics, Management and Planning with less than 10% in both

cases. Another important aspect is the fields where the FA approach retrieves more publications than the FWF records. These fields are the Medical Sciences, Chemistry, Physics and Astronomy, Earth and Environmental sciences, Mathematics, Statistics and Computer Sciences or Engineering Sciences. The fields where the FA analysis retrieves fewer publications are particularly Social Sciences and Economics, Management and Planning where the FA approach brings less than 30% of the publications funded by the FWF. Finally, if we focus on the percentage of FA 'forgetfulness' (i.e. the percentage of publications that are in the FWF system but they don't carry any FA to the FWF), we can see how this is particularly high in the Social and Economic sciences (particularly in the second) being higher than 50% in both cases. Interestingly, we can see a kind of negative relationship between the share of FA across disciplines and the share of 'forgetfulness' of acknowledgements in the fields, this probably suggesting that in those fields where funding is less frequent, the 'culture' and tradition of acknowledging funders is less established and therefore the authors 'forget' more frequently the acknowledgements of their funders, thus also indicating that the lower share of FA observed in this study and in other studies for these fields (e.g. Costas & van Leeuwen, 2012) could be relatively underrepresented.

Discussion of the results and further research

This paper presents a cross-field case study on the presence of funding acknowledgements among the publications funded by the Austrian Science Foundation – FWF using the new information gathered by Thomson Reuters Web of Science from the funding acknowledgements of scientific publications. This study presents different important novel elements: in the first place, to the best of our knowledge there are no other studies that combine the analysis of records collected and verified by a research funding organization, together with an extensive FA analysis. Secondly, there are no studies that have checked the impact of these publications and how the results can differ depending on the data collection method; and thirdly, there are no studies that have performed such an analysis across fields as in this case.

The development of this study already signals one of the most important challenges that this new FA analysis presents, namely, the great variance in the names of the funding organizations (. Another important result of this study is that more than 50% of the publications funded by the FWF are not included in the FWF records. The main explanation for this observation is the situation that the FWF gathers the publication information from the final reports of the projects. Thus, those publications that are published after the finalization of the project are lost from these records. Therefore, this result is informative of the important amount of publications (and citations) that could be lost in the records of funding agencies as we can argue that this problem is not only for the FWF but likely for most research funding agencies all around the world. Thus, it can be suggested that the analytical combination of FA analysis with the records maintained in the systems of funding agencies (or in other words, the combination of the recipient

and funder information) is the best approach in order to grasp the most complete picture of the activities and influence of these organizations.
As a conclusion, it can be suggested that FA analysis has a strong potential for the study of funding organizations and how they are shaping and influencing the scientific landscape. One open question that however still remains is about the share of publications for which authors should acknowledge funding but they don't do it at all. This question should be addressed in the future in order to determine the real scope of FA analysis.

Acknowledgments

The authors acknowledge all the valuable comments and suggestions by Falk Reckling, Ralph Reimann and Rudi Novak from the Austrian FWF.

References

Costas, R. & van Leeuwen, T.N. (2012). Approaching the "reward triangle": general analysis of the presence of funding acknowledgments and "peer interactive communication" in scientific publications. *Journal of the American Society for Information Science and Technology*, 63(8), 1647-1661.

Van Wijk, E. & Costas-Comesaña, R. (2012). *Bibliometric study of FWF Austrian Science Fund 2001-2010/11*. Leiden: CWTS-Leiden University.

Hornbostel, S. (2001). Third party funding of German Universities. An indication of research activity? *Scientometrics*, 50(3): 523-537.

Rigby, J. (2011). Systematic grant and funding body acknowledgement data for publications: new dimensions and new controversies for research policy and evaluation. Research Evaluation, 20(5), 365-375.

Rigby, J. (2013). Looking for the impact of peer review: does count of funding acknowledgements really predict research impact? *Scientometrics*, 94, 57-73.

Tiew, W.S. & Sen, B.K. (2002). Acknowledgement patterns in research articles: a bibliometric study based on Journal of Natural Rubber Research 1986-1997. Malaysian Journal of Library & Information Science, 7(1), 43-56.

Waltman, L., van Eck, N.J., van Leeuwen, T.N., Visser, M.S. & van Raan, A.F.J. (2011a). Towards a new crown indicator: some theoretical considerations. *Journal of Informetrics*, 5(1), 37-47.

Waltman, L., van Eck, N.J., van Leeuwen, T.N., Visser, M.S. & van Raan, A.F.J. (2011b). Towards a new crown indicator: an empirical analysis. *Scientometrics*, 87(3), 467-481.

PREDICTING AND RECOMMENDING POTENTIAL RESEARCH COLLABORATIONS

Raf Guns[1] and Ronald Rousseau[2]

[1] *raf.guns@ua.ac.be*
University of Antwerp, Institute for Education and Information Sciences, IBW,
Venusstraat 35, B-2000 Antwerp, Belgium

[2] *ronald.rousseau@khbo.be*
University of Antwerp, Institute for Education and Information Sciences, IBW,
Venusstraat 35, B-2000 Antwerp, Belgium
VIVES (Association KU Leuven), Faculty of Engineering Technology,
Zeedijk 101, B-8400 Oostende, Belgium
KU Leuven, B-3000 Leuven, Belgium

Abstract
We study research collaborations between cities in Africa, the Middle East and South-Asia, focusing on the topics of malaria and tuberculosis. For this investigation we introduce a method to predict or recommend high-potential future (i.e., not yet realized) collaborations. The proposed method is based on link prediction techniques. A weighted network of co-authorships at the city level is constructed. Next, we calculate scores for each node pair according to three different measures: weighted Katz, rooted PageRank, and SimRank. The resulting scores can be interpreted as indicative of the likelihood of future linkage for the given node pair. A high score for two nodes that are not linked in the network is then treated as a recommendation for future collaboration.
Results suggest that – of the three measures studied – the weighted Katz method leads to the most accurate predictions. Cities that often take part in new intercity collaborations are referred to as facilitator cities.

Conference Topic
Collaboration Studies and Network Analysis (Topic 6)

Introduction
Research collaboration is an important topic in informetrics. Collaboration has the potential of saving costs and diffusing insights and ideas between partners, a point also made in (Liu, Rousseau & Guns, 2013). Hence, the advantages of collaboration are especially attractive to institutes in those regions or countries that do not yet belong to the 'rich and famous' in science. While it may seem most attractive to collaborate with wealthier regions, there are several advantages when collaborating among developing nations, such as the establishment of local centres of excellence and a greater awareness among partners of the needs and problems common to developing nations (Boshoff, 2010).

In this article, we study research collaboration between cities in Africa, the Middle East, and South-Asia. We construct co-authorship networks among these cities within the research fields of malaria and tuberculosis during three consecutive, five-year time periods: 1997–2001, 2002–2006, and 2007–2011. Our aim is to develop a methodology for recommending potentially fruitful collaborations, using link prediction techniques. By comparing the recommendations for the first period with actual collaborations in the second, and recommendations for the second period with actual collaborations in the third, we can evaluate the quality of our recommendations.
In the next section, we discuss how the data has been collected. Subsequently, we discuss the extraction of the collaboration networks. We then explain our link prediction approach and highlight the recommended collaborations. The final section contains the conclusions.

Framework for data collection

Cities located in the following countries (referred to as the target countries) are included if they have contributions in the field under study:

- all African countries;
- all countries in the Middle East, except for Israel and Turkey (considered to be more European oriented);
- countries in South-Asia, that is, all Asian countries excluding countries that belong to the former Soviet Republic, Mongolia, China, North and South Korea, Taiwan and Japan.

As we are interested in collaborations between African and/or South-Asian cities on specific topics, we were seeking topics that were not entirely dominated by Western countries on the one hand, and not too specific to a certain country or region (The STIMULATE-6 Group, 2007) on the other. After some experimentation, we settled upon two diseases as topics: *malaria* and *tuberculosis*.
The data were collected from Thomson Reuters' Web of Science (WoS) on October 26 and November 21, 2012. We searched for all publications published in the three five-year periods (1997–2001, 2002–2006, 2007–2011) with at least one address in one of the target countries. These sets were then restricted to the two topics. Results are summarized in Table 7.

Table 7. Numbers of publications for each topic and period

Topic	Number of publications		
	1997–2001	2002–2006	2007–2011
malaria	2,622	4,671	7,901
tuberculosis	2,369	3,830	7,832

Table 8. Number of cities in the data

Topic	Number of cities (African and South-Asian / other)		
	1997–2001	2002–2006	2007–2011
malaria	400 / 361	601 / 587	904 / 883
tuberculosis	351 / 270	482 / 468	831 / 777

Methods

After exporting the search results from the WoS, we extracted a weighted network of co-authorship between cities as follows (for both topics and for each time period). For each publication, the city of each author's (primary) affiliation was recorded. A script was written to extract the city automatically. However, because of the large variety of address formats and inconsistencies in the data, all results were manually checked and corrected where necessary. Table 12 summarizes the results.

Subsequently a network was created whose node set consists of all cities encountered. All cities that co-occur on a single publication are then linked in the network. The weight of the link between cities A and B is the number of publications with authors from A and B. Because our analysis is on the level of cities rather than individuals, we have not taken into account the number of authors from a city on a single publication. For instance, a publication with five authors from city A and three from city B is treated the same as a paper with one author from A and one from B.

Some publications in our data have co-authors from cities outside the set of target countries (see 'other' in Table 2). Therefore, we decided to create two networks for each topic: a network including these external cities – the full network – and a network excluding them – the restricted network. In total, this procedure led to twelve different networks: a full and a restricted network for each of the two topics, and this for each of the three periods.

Collaboration network structure

Using VOSViewer (Van Eck & Waltman, 2007, 2010) we obtained twelve visualizations of our data: one for each network.

The malaria networks can be described as follows. In the full view (period 1997–2001) we can first see a dense main cluster dominated by Oxford, London and Bangkok; Indian cities (New Delhi) have a peripheral position. During the period 2002–2006 the main cluster is dominated by London, Bangkok and Nairobi. Indian cities have moved closer to the main cluster. Finally, during the period 2007–2011 we have a strong main cluster, including Indian cities, and dominated by London and Oxford. When considering the restricted networks the 1997–2001 view is rather scattered with centres in Nairobi and Bangkok, with some Vietnamese cities between these two centres; Indian cities are situated far away from these clusters. During the period 2002–2006 the Vietnamese cluster has almost merged with the Thai one. Finally during the period 2007–2011 there is a clear African cluster (Nairobi, Dakar, Cape Town) and an Asian one (Bangkok,

Mae Sot, New Delhi) as can be seen in Figure 11. Moreover an Iranian group of cities becomes visible on the periphery.

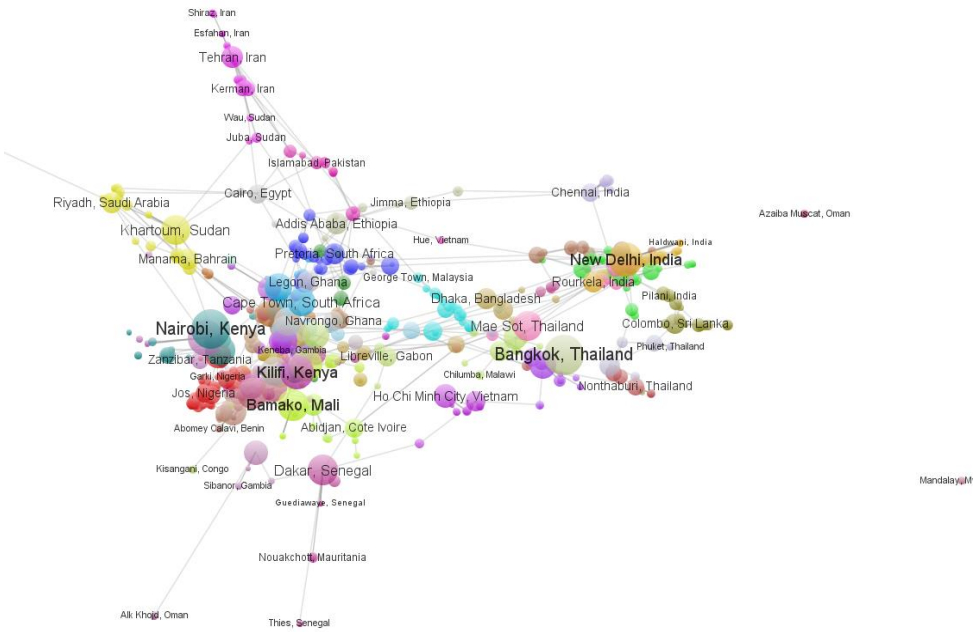


Figure 11. Collaboration network for malaria (restricted view, 2007–2011)

As to the tuberculosis networks, the 1997–2001 full view shows a group of centres around London, Geneva, Atlanta, Paris and Johannesburg. These are situated rather close to one another. During the period 2002–2006 these groups have formed a main cluster where we see London, Geneva, Paris, New Delhi and Oxford; Dhaka (Bangladesh) is clearly visible above this main cluster, while Antwerp and Brussels (Belgium) are situated in the very centre of this figure (Figure 12). In the 2007–2011 view we again have several clusters situated close to one another. The largest, central, one contains London, Paris, Geneva, Cape Town, Kampala and Liverpool; close to this main cluster we have an Indian cluster around New Delhi and Chennai; we further have clusters around Taipei and around Tehran. The 1997–2001 restricted network contains several scattered clusters around the following centres: South-Africa (Cape Town, Tygerberg, Johannesberg), Chennai-Pune, another Indian one around New Delhi, Bangalore and including Bangkok, and finally one around Addis Ababa (Ethiopia). The 2002–2006 view is very linear with centres around New Delhi, Hanoi, Bangkok and Cape Town (and other South African cities) including Dakar (Senegal). Finally the restricted 2007–2011 view contains a large cluster around Cape Town

(and South African cities) and including Addis Ababa. Moreover we see an Indian cluster, a Thai one and an Iranian one on the periphery.

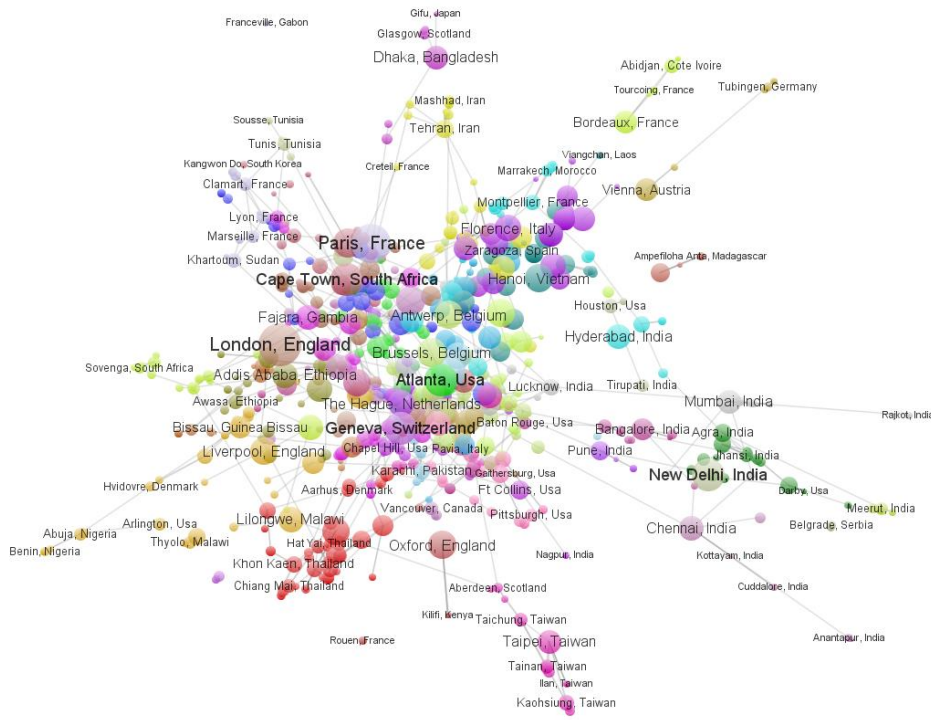


Figure 12. Collaboration network for tuberculosis (full view, 2002–2006)

In summary, the following observations pertain to both topics. The full networks are mainly dominated by Western cities, although some larger African or Asian cities are also able to occupy a central position. There appears to be at least a mild form of geographical bias – e.g., Asian cities mainly collaborating with other Asian cities – but the effect is modest: we also found several cases of intense international and intercontinental collaboration. Some countries, such as India and Iran, are more likely to form separate clusters. This observation corresponds with the results of Glänzel and Gupta (2008) who found that India has relatively few research collaborations with other countries.

Link prediction for recommendation

Since we are interested in opportunities for future collaboration, we focus on cities that do not yet collaborate in a given time period. There are many possible methods for determining which future collaborations are the most promising.

Here, we focus on the information that is already present in the city collaboration network, without relying on any other data source. We start from the assumption that a collaboration should be recommended if (a) the two cities do not yet collaborate, and (b) the two cities are similar or related. To determine the similarity or relatedness of cities, we take a link prediction approach. We try to determine a relatedness score *W* for each node pair on the basis of the current network. Singling out those pairs that are currently unlinked (condition a) and sorting them in decreasing order of *W* (condition b) yields a list of the most promising future collaborations.
A formula that results in a relatedness score *W* is called a predictor. We have used three predictors that had good performance in previous research (Guns, 2011, 2012; Liben-Nowell & Kleinberg, 2007): weighted Katz, rooted PageRank, and SimRank.

Weighted Katz predictor

Before we define the (weighted) Katz predictor (Katz, 1953) we explain the used terminology. A *walk* is a sequence of nodes $v_1, v_2, \dots, v_m$, such that each node pair v_i, v_{i+1} in the sequence is connected by a link. There are no further restrictions on walks. A *multigraph* is a graph allowed to have multiple links between two nodes. Different links between two nodes also constitute different walks, i.e. the number of walks $v_1, v_2, \dots, v_m$ in a multigraph is equal to $\prod_{i=1}^{m-1} N(v_i, v_{i+1})$, where $N(v_i, v_{i+1})$ denotes the number of links between v_i and v_{i+1}.

Weighted Katz measure: definition

The weighted Katz measure can best be described in the context of a multigraph. Let A denote the (full) adjacency matrix of the multigraph M. The element a_{ij} is equal to the number of links between v_i and v_j or 0 if no link is present. Each element $a_{ij}^{(k)}$ of A^k (the k-th power of A) has a value equal to the number of walks in M with length k from v_i to v_j (Wasserman & Faust, 1994, p. 159). The weighted Katz predictor is then defined as:

$$W(v_i, v_j) = \sum_{k=1}^{\infty} \beta^k a_{ij}^{(k)} \qquad (1)$$

where β is a parameter between 0 and 1. This parameter represents the "probability of effectiveness of a single link". Thus, each path with length k has a probability β^k of effectiveness. As $0 < \beta < 1$ higher powers become smaller and smaller so that the influence of nodes further away decreases fast.

Rooted PageRank

The other two predictors are inspired by Google's PageRank (and hence indirectly by the Pinski-Narin citation influence methodology (1976)). The intuition behind rooted PageRank (Liben-Nowell & Kleinberg, 2007) is best explained from the perspective of a random walker. The random walker starts at a fixed node v, called the root node. At each step, the walker moves along a link to a neighbour of the current node. Contrary to ordinary PageRank, rooted PageRank does not allow random 'teleportation' but only allows teleportation back to the root node v. This form of teleportation occurs with probability $1-\alpha$ (where $0 < \alpha < 1$). High α values tend to favour the well-connected nodes in the network (with high classic PageRank scores), especially in relatively small networks such as ours. On the other hand, setting α too low reduces the advantage of a PageRank-like predictor. Essentially, rooted PageRank is a specific form of so-called personalized PageRank (Langville & Meyer, 2005). The resulting scores can be interpreted as a measure of each node's relatedness to the root node. The highest scoring node is typically the root node itself.

SimRank

SimRank is a measure of how similar two nodes in a network are, originally proposed by Jeh and Widom (2002) and further elaborated by Antonellis, Molina, and Chang (2008). The SimRank thesis can be summarized as: *Objects that link to similar objects are similar themselves*. Note the recursive nature of the thesis, – to assess the similarity of a node pair, we need to have an estimate of the similarity of the nodes that they link to. The starting point of a SimRank computation is the assumption that an object is maximally similar to itself: $Sim(a,a) = 1$. One can then calculate the SimRank score of each node pair iteratively, until the changes drop below a given threshold value. The SimRank formula is:

$$W(u,v) = \frac{c}{|N_u| \cdot |N_v|} \sum_{p \in N_u} \sum_{q \in N_v} W(p,q) \qquad (2)$$

where N_v denotes the neighbourhood of v (the set of nodes adjacent to v), and $|N_v|$ the number of neighbours of v. In case of isolate nodes, the above formula would lead to a division by zero, which can be avoided by adding 1 to the denominator. Since our data contains no isolates, this is not necessary.
In (2), c $(0 < c < 1)$ is the 'decay factor' that determines how quickly similarities decrease. If, for example, cities x and y both collaborate with z, then c determines the certainty with which we can state that x and y are similar. Lower c values also result in lower values for $W(x,y)$.

Results

We predicted collaborations between research institutes situated in different cities based on relatedness scores as explained above. The parameters for each predictor were set to the values shown in Table 9.

Since we are interested in recommending high-potential collaborations, it makes sense to restrict our analysis to the top predictions. Concretely, for each method we drew a list of the twenty unlinked node pairs with the highest relatedness score. These can be considered as our recommendations according to that method. To evaluate the quality of the recommendations, we applied this procedure to the networks from the periods 1997–2001 and 2002–2006. We consider a recommendation successful if it actually took place in the following period. In this way, we determined the success rate, the fraction of realized collaborations. Results are shown in Figure 13.

Table 9. Values chosen for predictor parameters

Predictor and parameter	Value
Weighted Katz: β	0.001
Rooted PageRank: α	0.4
Simrank: c	0.3

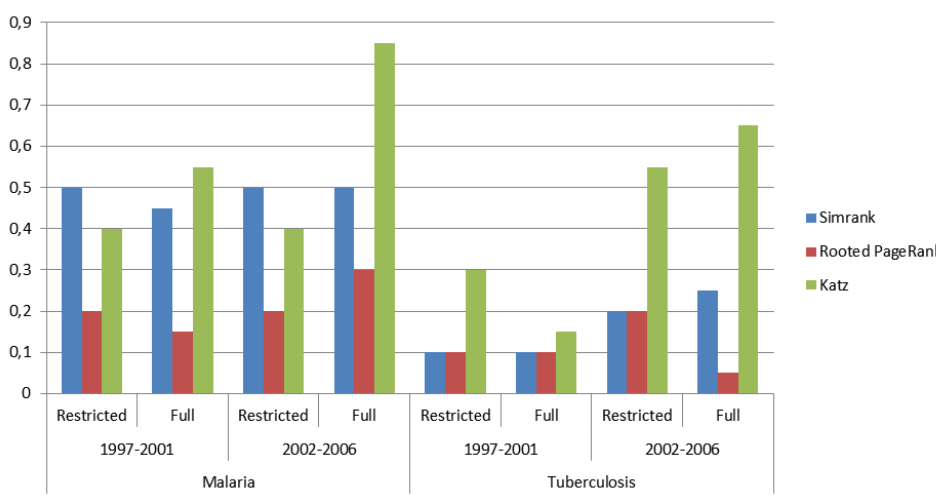


Figure 13. Results of three predictors in two medical research fields for two periods. The horizontal axis refers to the success rate

Figure 13 clearly shows that overall the weighted Katz predictor performs best, followed by SimRank. Rooted PageRank is the weakest predictor in our study. Predictions based on the full network are generally (but not always) better than those based on the restricted network. Since the former contains more information than the latter, this is not unexpected. Indeed, if two target cities collaborate with

a Western city, they may eventually end up collaborating directly, but this can only be inferred from the full network.
Predictions based on the second period are generally better than those based on the first one. This is most likely because the network grew larger and hence contained more information. Successful predictions often involved South-African or Thai cities (e.g., Bangkok, Mae Sot, Johannesburg). Such cities could be called facilitator cities. They play a central role in weaving the fabric of international collaboration. In our opinion, two findings in particular bear testimony of the potential of our approach. First, we obtained high success rates, especially for the weighted Katz predictor. Second, among the correctly predicted collaborations we have several ones involving a city in Asia as well as a city in Africa. This illustrates that the method outlined in this paper is capable of making realistic but non-trivial recommendations.

Conclusions

This investigation shows that it is possible – at least to some extent – to predict collaborations. Yet, we can also consider our predictions as recommendations for future research and as long as these recommendations are not actually made and tried out, it is possible that some institutes just have missed some excellent potential collaborators. Of course, this aspect cannot be evaluated.
Although it seems that the larger the network, the better the predictions, it is obvious that there is an upper limit on the size or density of the network. After a while only highly improbable new collaborations can be predicted.
By focussing on cities and regions this article contributes to the emerging subfield of spatial or regional scientometrics (Frenken, Hardeman & Hoekman, 2009).

References

Antonellis, I., Molina, H.G., & Chang, C.C. (2008). Simrank++: query rewriting through link analysis of the click graph. *Proceedings of the VLDB Endowment,* 1(1), 408–421.

Boshoff, N. (2010). South–South research collaboration of countries in the Southern African Development Community (SADC). *Scientometrics*, 84(2), 481–503.

Frenken, K., Hardeman, S., & Hoekman, J. (2009). Spatial scientometrics. Towards a cumulative research program. *Journal of Informetrics*, 3(3), 222-232.

Glänzel, W., & Gupta, B.M. (2008). Science in India. A bibliometric study of national research performance in 1991-2006. *ISSI Newsletter*, 4(3), 42–48.

Guns, R. (2011). Bipartite networks for link prediction: can they improve prediction performance? In E. Noyons, P. Ngulube, & J. Leta (Eds.), *Proceedings of the ISSI 2011 Conference* (pp. 249–260). Durban: ISSI, Leiden University, University of Zululand.

Guns, R. (2012). *Missing links: Predicting interactions based on a multi-relational network structure with applications in informetrics*. Doctoral dissertation, Antwerp University.

Jeh, G., & Widom, J. (2002). SimRank: a measure of structural-context similarity. In *KDD '02: Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 538–543). New York: ACM.

Katz, L. (1953). A new status index derived from sociometric analysis. *Psychometrika*, 18(1), 39–43.

Langville, A.N., & Meyer, C.D. (2005). A survey of eigenvector methods for web information retrieval. *SIAM Review*, 47(1), 135–161.

Liben-Nowell, D., & Kleinberg, J. (2007). The link-prediction problem for social networks. *Journal of the American Society for Information Science and Technology*, 58(7), 1019–1031.

Liu, YX., Rousseau, R., & Guns, R. (2013). A layered framework to study collaboration as a form of knowledge sharing and diffusion. *Journal of Informetrics* (to appear).

Pinski, G. & Narin, F. (1976). Citation influence for journal aggregates of scientific publications: Theory with application to the literature of physics. *Information Processing & Management*, 12(5), 297–312.

The STIMULATE-6 Group (2007). The Hirsch index applied to topics of interest to developing countries. *First Monday*, 12(2). http://www.firstmonday.org/issues/issue12_2/stimulate/

Van Eck, N.J., & Waltman, L. (2007). VOS: a new method for visualizing similarities between objects. In H.-J. Lenz, & R. Decker (Eds.), *Advances in Data Analysis: Proceedings of the 30th Annual Conference of the German Classification Society* (pp. 299-306). Springer.

Van Eck, N.J., & Waltman, L. (2010). Software survey: VOSviewer, a computer program for bibliometric mapping. *Scientometrics*, 84(2), 523–538.

Wasserman, S., & Faust, K. (1994). *Social Network Analysis: Methods and Applications*. Cambridge: University Press.

PUBLICATION BIAS IN MEDICAL RESEARCH: ISSUES AND COMMUNITIES

Edgar Schiebel[1] and Maria-Elisabeth Züger[2]

[1] *edgar.schiebel@ait.ac.at,* [2]*maria-elisabeth.zueger@ait.ac.at*
AIT Austrian Institute of Technology GmbH, Donau City Str 1, A-1220 Vienna (Austria)

Abstract
Publication bias is a broadly discussed phenomenon related to the reporting of the outcome of clinical studies. As a part of the UNCOVER project this work aimed at the identification of members of the research community as stakeholders for interviews and workshops how to overcome publication bias.
For this objective relevant literature was analysed by the following bibliometric approaches: networks of co-authorships and affiliated institutions, co-citation analysis and bibliographic coupling over a twenty-year timespan. Research communities were mapped and examined and research issues where identified by applying bibliographic coupling and co-citation maps.
The analysis showed a high dominance of publications from evidence based medicine like systematic reviews and meta-analysis performed for different medical topics. Most authors used and cited previous research, findings and methods how to proceed with publications bias. They are to be seen as experienced "users and applicants" stakeholder group.
A second dominant group of publications is related to research on publication bias from different aspects: publications and data for systematic reviews, adequacy of databases, publication of negative results, registration of clinical trials, outcome reporting, protocols of clinical trials, sponsorship bias, role of editors, ethic committees, guidelines for systematic reviews, regulation of clinical trials and methods for meta-analysis. Stakeholders were selected on the basis of research issues and affiliated organizations.

Conference Topic
Science Policy and Research Evaluation: Quantitative and Qualitative Approaches – (Topic 3); Visualisation and Science Mapping: Tools, Methods and Applications (Topic 8)

Introduction
It has already been discussed, that non-publication of negative results is a challenge in science (Gumpenberger, C., Gorraiz, J., Wieland, M., Roche, I., Schiebel, E., Besagni, D., Francois, C., 2012 and Roche, I., Francois, C. Gumpenberger, C., Gorraiz, J., Wieland, M., Schiebel, E., 2012). More efficiency, more progress and more transparency could be achieved, if not only positive outcomes but also non successful approaches were published. In evidence based medicine publication bias is a well-known phenomenon (Song F, Parekh S,

Hooper L, Loke YK, Ryder J, Sutton AJ, HIng C, Kwok CS, Pang C, Harvey I., 2010).

The framework for the presented bibliometric analysis is formed by the UNCOVER project: Evaluation and development of measures to uncover and overcome bias due to non-publication of clinical trials –. The UNCOVER project is a direct contribution to overcome non-publication of clinical studies that have been designed and executed as randomized controlled trials (RCTs). The issues of the publication bias are treated with quantitative, qualitative and participatory means in an interdisciplinary approach (Züger, ME., Holste, D. Schiebel, E., 2013). It is framed in terms of evidence-based medicine and system's theory. The core of the system approach is twofold. First we identify, map and link relevant stakeholders and opinion leaders like study registries, researchers, editors of journals, funding bodies, regulators, and industry on an international and global level, and secondly possible measures (law, regulations, policies, practices, guidelines, methods, tools) are identified to overcome bias. Current measures substantiated by own experience ("inside-out") are identified by a systematic review. Experts from international methods groups ("outside-in") in the field of systematic reviews and meta-analyses are engaged. Measures in terms of experiences, own strategies and existing conflict of interests are reflected by personal interviews with editors and other stakeholders based on stakeholder mapping/analysis. Software solutions for the demonstration and treatment of unpublished studies on statistical meta-analyses are developed. Recommendations imply the implementation of feasible measures and milestones, as well as open gaps addressed by new research, to overcome non-publication.

UNCOVER thus both provides viable solutions for the publication bias for better allocation efficiency of medicinal and health related research funds, and develops methodologies for future bias research efforts.

In this paper we present a bibliometric analysis of scientific literature on publication bias as a general issue: Do we have a scientific discussion about publication bias in science and what are the concerned disciplines; Are scientists aware and do they mention and examine non-publication of negative results that also plays a role for the quality of published work and have a consequence in practical use of research results? We assume that non-publication of negative results or inflation of positive ones has many consequences in medical research, especially in reporting the outcome of clinical trials.

We present a literature survey performed by relational bibliometric approaches to identify research groups, issues and disciplines with research on publication bias.

Method and Data

The bibliometric analysis was aimed at the identification of members of the research community and research issues in the field of publication bias assisted by a quantitative bibliometric approach.

To this end, bibliographic data (e.g., title, authors, institution, country, abstract, keywords, references) of the relevant literature using the search phrases

"publication bias", "citation bias", "language bias", "location bias", "reference bias", and "reporting bias" was obtained from the ISI Web of Knowledge.
Based on 3,891 publications over the time span 1990 to 2012 (partly), bibliometric analysis was conducted on co-authorships, affiliated institutions, and bibliographic coupling. Relationships between authors, institutions were mapped and examined with a network analysis. Research issues were identified by applying bibliographic coupling.
Bibliographic coupling is of growing interest to subdivide research fields in research issues, sometimes called research fronts. Basic work about the identification of research fronts by bibliographic coupling and knowledge bases by co-citation analysis was published by Price (1965), Kessler (1965), Chen & Morris (2003) and more actual research was performed by Shibata et al (2009) and Boyack & Klavans, (2010) just to cite some of the growing amount of publications in this issue. Spring models and multidimensional scaling are used to map publications in a two or more dimensional space, see for example Kopcsa, A. and Schiebel, E. (1998). Advantages are the visualisation, the representation in two or three dimensions maps and the visibility of relative positions of agglomerations. Schiebel (2012) introduced a plot of link-weighted local densities of publications or references that was used in this work to identify research issues on publication bias.

Results

The results section consists of different descriptive statistics about titles of sources (journals, proceedings, etc.), time series and countries data with regard to the number of publications. It gives information about discipline specific journals, timeliness and geographic engagement.
Table 1 lists journals as a percentage of all journals for of the 3,891 publications. The first 21 Journals sum up to nineteen percent of all publications.
We have a high dominance of publications in the Cochrane Database of Systematic Reviews, see The Cochrane Collaboration, (2013). The British Medical Journals is the second most important source for publications relevant to publication bias. Other journals are the Journal of Clinical Epidemiology, JAMA Journal of the American Medical Association and Annals of International Medicine just to cite media with more than 50 publications. All sources are from the medical discipline. As the ISI Web of Knowledge database covers all scientific disciplines it can be concluded, that publication bias is primarily an issue in medical research.
A small number of publications dates back almost two decades, to the year 1990 – the starting point of our analysis (cf. Table 2). Yet it took more than the first decade (about 14 years) to attain a remarkable increase in the number of publications in this field. Since then, the number of publications is monotonically increasing with an approximately constant growth rate. In the last two years there are indications for further acceleration of the growth rate (and the numbers for 2012 tend to rise further). The growth per year is indicative of the increasing

research on publication bias from different perspectives like outcome reporting, registration of trials, ethic issues, role of editors, guidelines for performing clinical trials reporting and the increase of the number of systematic reviews and meta-analyzes on different medical topics. It reflects the growing research activities in evidence based medicine and awareness of publication bias.

Table 1: List of the first 21 journals sorted by descendant number of publications

	Source titles	*Number of Publ.*	*% of 3,891*
1.	COCHRANE DATABASE OF SYSTEMATIC REVIEWS	101	2.60
2.	BRITISH MEDICAL JOURNAL	83	2.13
3.	JOURNAL OF CLINICAL EPIDEMIOLOGY	78	2.01
4.	JAMA JOURNAL OF THE AMERICAN MEDICAL ASSOCIATION	57	1.47
5.	ANNALS OF INTERNAL MEDICINE	53	1.36
6.	AMERICAN JOURNAL OF EPIDEMIOLOGY	41	1.05
7.	LANCET	36	0.93
8.	STATISTICS IN MEDICINE	35	0.90
9.	PLOS ONE	34	0.87
10.	INTERNATIONAL JOURNAL OF EPIDEMIOLOGY	24	0.62
11.	EPIDEMIOLOGY	23	0.59
12.	AMERICAN JOURNAL OF CLINICAL NUTRITION	20	0.51
13.	PLOS MEDICINE	20	0.51
14.	STROKE	20	0.51
15.	AMERICAN JOURNAL OF GASTROENTEROLOGY	19	0.49
16.	MOLECULAR BIOLOGY REPORTS	19	0.49
17.	ARCHIVES OF INTERNAL MEDICINE	17	0.44
18.	CANADIAN MEDICAL ASSOCIATION JOURNAL	17	0.44
19.	CURRENT MEDICAL RESEARCH AND OPINION	17	0.44
20.	EUROPEAN JOURNAL OF CANCER	16	0.41
21.	JOURNAL OF CLINICAL ONCOLOGY	16	0.41

The field is headed by North America and dominated by the United States (with 1,480 publications in the whole period), where we have the highest publication activity. A large number of European countries are listed as the address of authors and their affiliated institutions. England (with 760 publications) is leading the statistics of European countries. Positioned on the fourth place (with 346 publications), China plays a key role, too, but is not dominating like it does in many engineering domains.

Keywords were derived from three record-fields: the publication title (TI); the author key-words (DE) and the Web of Knowledge keywords (ID). The most frequent terms indicate that publication bias is strongly connected to meta-analysis systematic reviews and clinical trials and not to other areas of science disciplines.

Table 2: Number of publications related to publication bias per year.

Publication Year	*Number of Publ.*	*% of 3,891*	*Trend line*
1990	4	0.10	
1991	26	0.67	
1992	42	1.08	
1993	43	1.11	
1994	52	1.34	
1995	52	1.34	
1996	65	1.67	
1997	78	2.01	
1998	93	2.39	
1999	108	2.78	
2000	98	2.52	
2001	105	2.70	
2002	119	3.06	
2003	124	3.19	
2004	163	4.19	
2005	218	5.60	
2006	244	6.27	
2007	276	7.09	
2008	353	9.07	
2009	400	10.28	
2010	422	10.85	
2011	531	13.65	
2012 (January-May)	275	7.07	

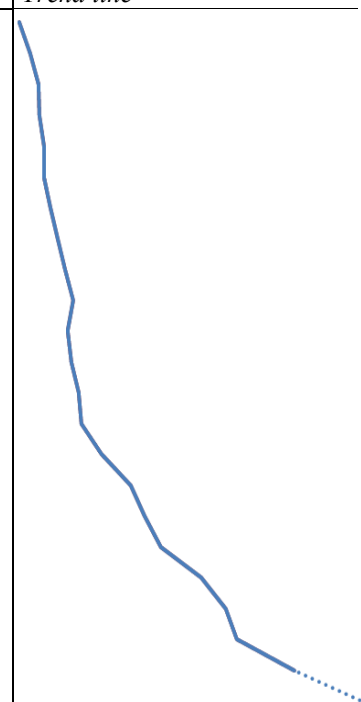


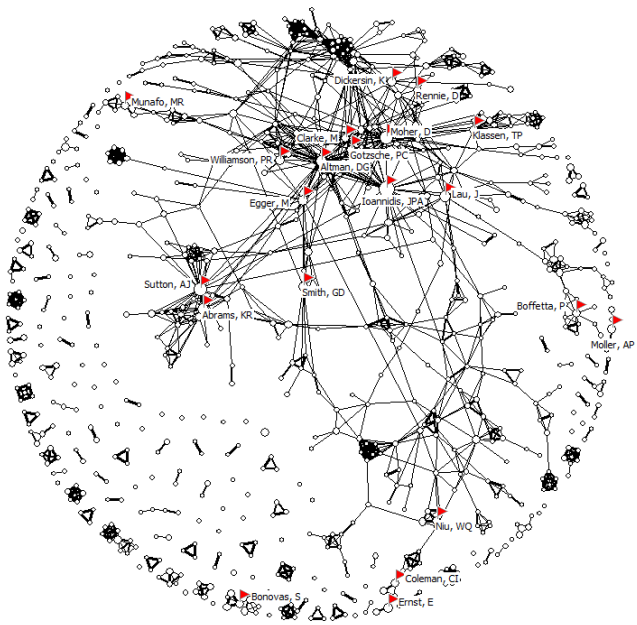
Figure 1 shows the co-authorship network with more than 3 publications per author. The top 21 authors (ranked by the number of publications) are marked in the graph. The network is dominated by a giant sub-network, which consists of a highly inter-linked core, with many of the highest active researchers and with connections to various working groups via authors in a network role as brokers (central position in a sub-network) or bridges (connecting one or more sub-networks). Almost 60% of authors mapped in the graph belong to this dominant network component.

In addition, the graph shows a number of smaller components with a size of in-between 2 to 5 authors with more than 2 publications.

The structure of the author network is nothing out of the ordinary compared to other re-search fields. Although it is unusual that so many of the top 21 authors are linked instead of having their own work groups connected indirectly via brokers and bridges.

Sub-networks were defined as a group of authors which are only connected to each other. As shown in figure 1 there is one sub-network of 442 authors. To make further analysis possible this sub-network – with the representative Ioannidis, JPA – was subdivided into frequently co-operating working groups. This was achieved by hiding weak links which yields in several separated groups

of strongly connected authors. Authors of such a group are co-operating frequently and form a working group. Results are listed in tables 3 and 4.



Author	# Pub
Ioannidis, JPA	47
Moher, D	33
Sutton, AJ	33
Altman, DG	26
Abrams, KR	22
Lau, J	22
Egger, M	19
Dickersin, K	18
Clarke, M	15
Ernst, E	15
Smith, GD	15
Boffetta, P	14
Bonovas, S	14
Coleman, CI	14
Gotzsche, PC	14
Klassen, TP	14
Moller, AP	14
Munafo, MR	14
Niu, WQ	14
Rennie, D	14
Williamson, PR	14

Figure 1: Network of co-authorships. The top 21 authors (ranked by number of publications) are marked with flags and listed on the right-hand side; Nodes: authors, the size corresponds to the number of publications; Edges: Jaccard index of co-frequencies; Timespan of analysis: 1990 to 2012; Date of research: 06 2012; Total number of publications: 3,891; each author published at least 3 publications; Number of nodes: 754; Number of edges: 1,505.

The local density map of bibliographically coupled publications was used to identify research issues and key researchers related to publication bias. Research issues were identified by selecting an agglomeration of publications, listing keywords from the publications and reading titles and abstracts. The list of keywords was ranked the term frequency–inverse document frequency (TF-IDF) measure. The TF-IDF weights how often a keyword occurs in a chosen subgroup of the agglomeration relative to the overall occurrence. It achieves that very common and thus unspecific terms like “publication bias” have a low rank and specific keywords like “clinical trial” have a high rank.

The bibliographic coupling reveals two huge agglomerations of research activities.

The bigger one is formed by systematic reviews and meta-analyzes about different medical subjects like “myocardial infarction”, “blood preasure” or diabetis mellitus”. We called it “meta-studies about clinical topics”. It includes systematic reviews and meta-analyzes.

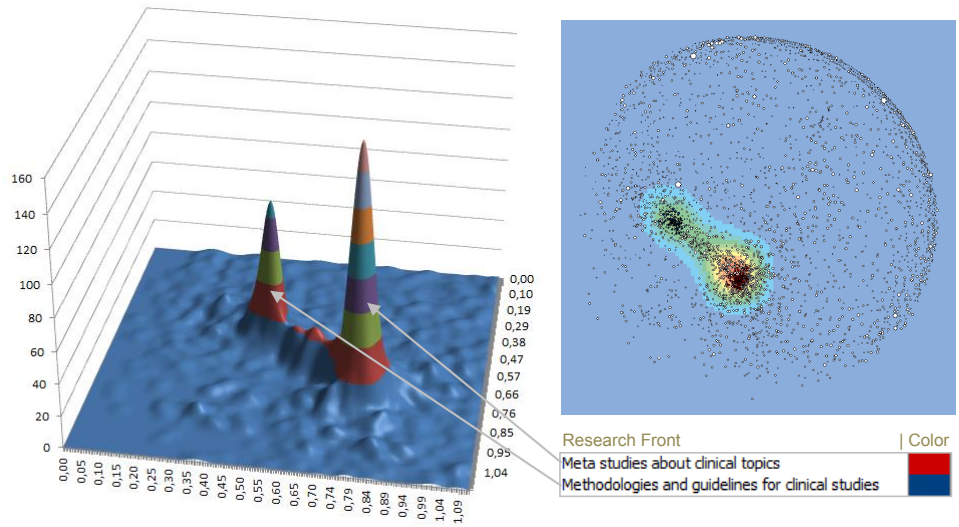


Figure 2: Local density of bibliographically coupled publications: 3D surface map [left] and 2D surface map [right]; dots are publications; total number of publications: 3.891; Number of edges: 1.177.507.

The second peak represents publications on research about publication bias. Published work reports on guidelines for clinical trials, mathematic and statistical methods for meta-analysis, registration of studies, reporting and research about different issues related to publication bias. The assigned name is: “methodologies and guidelines for clinical studies”.

The network of authors does not reflect the two clusters that were identified as agglomerations of bibliographically coupled publications. Generally spoken we have a sequence of links between authors who work on publication bias and authors who work on meta-analysis and systematic review for clinical subjects.

The task to identify stakeholders requires a two-fold interpretation of the author network based on the two identified clusters.

For each cluster authors were analyzed by using the following indicators: the number of publications, number of citations and recent publications. In addition to linear indicators, we studied relational information based on co-authorships, which reveals networks of research groups.

Table 3: Cluster "Meta studies about clinical topics": highest ranked authors by the number of publications in the cluster, including author network information and title of the highest cited publication in this cluster. Due to co-publications, different authors can have the same publication listed. Times cited as from June 2012.

Index	*Author*	*Author network*		*Author sub-networks*		*Publications in the two clusters*		
		Times cited (diameter)	No. of papers	Sub-network	Size of the sub-network	Highest cited publication	"Meta studies about clinical topics"	"Methodologies and guidelines for clinical studies"
1	Bagos, PG	11	12	Bonovas, S	8	Association between the plasminogen activator inhibitor-1 4G/5G polymorphism and venous thrombosis - A meta-analysis	10	
2	Nikolopoulos, GK	14	9	Bonovas, S	8	Association between the plasminogen activator inhibitor-1 4G/5G polymorphism and venous thrombosis - A meta-analysis	8	
3	Qin, LQ	6	9	Qin, LQ	5	Milk consumption is a risk factor for prostate cancer in Western countries: evidence from cohort studies	8	
4	Wang, J	10	12	Ioannidis, JPA	442	UCHL1 is a Parkinson's disease susceptibility gene	7	
5	Sutton, AJ	48	33	Ioannidis, JPA	442	Empirical assessment of effect of publication bias on meta-analyses	7	1
6	Dong, JY	1	7	Qin, LQ	5	Erectile Dysfunction and Risk of Cardiovascular Disease Meta-Analysis of Prospective Cohort Studies	7	
7	Xu, MQ	5	7	Ioannidis, JPA	442	Quantitative assessment of the effect of angiotensinogen gene Polymorphisms on the risk of coronary heart disease	6	
8	Mengoli, C	38	5	Cruciani, M	2	Use of PCR for diagnosis of invasive aspergillosis: systematic review and meta-analysis	5	
9	Cruciani, M	38	5	Cruciani, M	2	Use of PCR for diagnosis of invasive aspergillosis: systematic review and meta-analysis	5	
10	Abrams, KR	46	22	Ioannidis, JPA	442	A systematic review of molecular and biological tumor markers in neuroblastoma	5	

Table4: Cluster: "Methodologies and guidelines for clinical studies": highest ranked authors by the number of publications in the cluster, including author network information and title of the highest cited publication in this cluster. Due to co-publications, different authors can have the same publication listed. Times cited as from June 2012.

Index	*Author*	*Author network*		*Author sub-networks*		*Publications in the two clusters*		
		Times cited (diameter)	No. of papers	Sub-network	size of the sub-network	Highest cited publication	"Meta studies about clinical topics"	"Methodologies and guidelines for clinical studies"
1	Moher, D	128	33	Ioannidis, JPA	442	Does the inclusion of grey literature influence estimates of intervention effectiveness reported in meta-analyses?		9
2	Dickersin, K	178	18	Ioannidis, JPA	442	Systematic reviews - identifying relevant studies for systematic reviews		8
3	Decullier, E	44	5	Ioannidis, JPA	442	Systematic Review of the Empirical Evidence of Study Publication Bias and Outcome Reporting Bias		5
4	Tannock, IF	30	8	Tannock, IF	2	Factors associated with failure to publish large randomized trials presented at an oncology meeting		5
5	Ioannidis, JPA	78	47	Ioannidis, JPA	442	Systematic Review of the Empirical Evidence of Study Publication Bias and Outcome Reporting Bias	3	5
6	Rennie, D	217	14	Ioannidis, JPA	442	Publication bias in editorial decision making		4
7	Smith, GD	451	15	Ioannidis, JPA	442	Sifting the evidence - what's wrong with significance tests?		4
8	Egger, M	344	19	Ioannidis, JPA	442	Uses and abuses of meta-analysis		4
9	Cook, DJ	361	9	Ioannidis, JPA	442	Users guides to the medical literature .6. how to use an overview		4
10	Bero, LA	133	8	Ioannidis, JPA	442	Pharmaceutical industry sponsorship and research outcome and quality: systematic review		4

Tables 3 and 4 represent the results of the analysis of the author network for the two identified research issues. It lists the first 21 authors (ranked by the number of published papers) and includes information on the author network and sub-networks. Additionally we included the publication that is highest cited.

Summary and Discussion

The aim of this work was to identify key researchers about "publication bias". The task was also aimed at mapping the published research activities in the field of publication bias, including thematic clustering.

In a first step, the selected set of publications was examined by means of descriptive statistics. As a second step relational maps for an authors network and research issues were drawn.

A small number of publications about 'publication bias' dates back almost two decades, to the year 1990 – the starting point of our analysis. Yet it took more than the first decade (about 14 years) to attain a remarkable increase in the number of publications in this field.

Since then, the number of publications is monotonically increasing with an approximately constant growth rate. In the last two years there are indications for further acceleration of the growth rate. The growth per year is indicative of the increasing research on publication bias from different perspectives like outcome reporting, registration of trials, ethic issues, role of editors, guidelines for performing clinical trials reporting and the increase of the number of systematic reviews on different medical topics. It reflects the growing research activities in evidence based medicine, awareness and methods for meta-analysis and systematic reviews.

The field is headed by North America and dominated by the United States (with 1,480 pub-lications), where we have the highest publication activity. A large number of European countries are listed as the address of authors and their affiliated institutions in the field of publication bias. England is leading the statistics of European countries. Positioned on the fourth place, China plays a key role, too, but is not dominating like it does in many engineering domains.

The author network showed a high level of co-publishing in the field of publication bias. Ranking authors by the number of publications, rank numbers 1–21 (the top 21) formed a large predominant cluster (or subnetwork). Interestingly, this large cluster displayed the network type of "brokers", i.e. authors within groups through a single or few links between groups. Given the accommodating information about authors (e.g. institution, country, or topic), network positions were used to optimize the selection of stakeholders for up-coming interviews and workshops in the UNCOVER project.

The map of bibliographically coupled publications showed two clusters: The bigger one is formed by publications about performed systematic reviews and meta-analysis about different medical subjects like "myocardial infarction", "blood preasure" or diabetis mellitus". We called it "meta-studies about clinical topics". It includes systematic reviews and meta-analyzes.

The second peak represents publications on research about publication bias. Publications are about guidelines for clinical trials, mathematic and statistical methods for meta-analysis, registration of studies, reporting and research about different issues related to publication bias. The assigned name was "methodologies and guidelines for clinical studies".
Key researches were identified by exploiting the network of authors combined with the results of the map of bibliographically coupled publications. We identified research communities and persons as stakeholders for publication bias and how to overcome it: improvement of performing clinical trials and reporting their out-come; study registration; sponsorship bias; editorial bias, statistical improvement of available trial results by meta-studies as well as systematic reviews.

Acknowledgments

This work is part of the project UNCOVER - Evaluation and development of measures to uncover and overcome bias due to non-publication of clinical trials; HEALTH.2011.4.1-2: Targeting publication bias; project number: 282574; with contributions from dirk Holste, AIT.

References

Boyack & Klavans, (2010). Co-Citation Analysis, Bibliographic Coupling, and Direct Citation: Which Citation Approach represents the Research Front Most Accurately?, JASIST 61(12): 2389-2404.

Chen, C., & Morris, S. (2003). Visualizing evolving networks: Minimum spanning trees versus pathfinder networks. Proceedings of IEEE Symposium on Information Visualization (pp 67-74), Seattle, WA: IEEE Computer Society Press

The Cochrane Collaboration, (2013), hompage: http://www.cochrane.org/cochrane-reviews/cochrane-database-systematic-reviews-numbers

Gumpenberger, C., Gorraiz, J., Wieland, M., Roche, I., Schiebel, E., Besagni, D., Francois, C. (2012) Exploring the bibliometric and semantic nature of negative results, Scientometrics 1-21

Kopcsa, A., Schiebel, E. (1998), Science and Technology Mapping: A New Iteration Model for Representing Multidimensional Relationships. Journal of the American Society for Information Science JASIS (1998), 49, 1, 7-17

Roche, I., Francois, C. Gumpenberger, C., Gorraiz, J., Wieland, M., Schiebel, E.: (2012): Análisis bibliométrico de la literatura secundaria sobre la publicación de resultados negativos; Congreso Internacional de Información - INFO 2012, La Habana, Cuba; 16.04.2012 - 20.04.2012.

Price, D.D. (1965) Networks of scientific papers. Science, 149, 510-515

Schiebel, E. (2012) Visualization of Research Fronts and Knowledge Bases by Three-Dimensional Areal Densities of Bibliographically Coupled Publications

and Co-Citations. Special Issue Scientometrics, DOI: 10.1007/s11192-012-0626-8
Shibata et al, (2009). Comparative Study on Methods of Detecting Research Fronts Using different Types of Citation. JASIST 60(3):571-580
Song F, Parekh S, Hooper L, Loke YK, Ryder J, Sutton AJ, HIng C, Kwok CS, Pang C, Harvey I. (2010). Dissemination and publication of research findings: an updated review of related biases. Health Technology Assessment; 14(8): 1-220
Züger, ME., Holste, D. Schiebel, E. (2013). Deliverable D3.1 (Part B) of the UNCOVER FP7-funded project under contract number 282574: Bibliometric analysis on publication bias. Publicly available under http://publicationbias.eu

QUANTITATIVE EVALUATION OF ALTERNATIVE FIELD NORMALIZATION PROCEDURES

Yunrong Li[1], Filippo Radicchi[2], Claudio Castellano[3] and Javier Ruiz-Castillo[1]

[1] *yli@eco.uc3m.es*
Universidad Carlos III de Madrid, Departamento de Economia, Madrid (Spain)

[2] *f.radicchi@gmail.com*
Universitat Rovira i Virgili, Department d'Enginyeria Quimica, Av. Paisos Catalans 26, 43007 Tarragona (Spain)

[3] *claudio.castellano@roma1.infn.it*
Istituto dei Sistemi Complessi (ISC-CNR), Via dei Taurini 19, 00185 Roma (Italy) and Dipartimento di Fisica, "Sapienza" Universita' di Roma, P.le A. Moro 2, 00185 Roma (Italy)

Abstract
The use of citation numbers for the assessment of research quality has become highly relevant in modern science. Although it is well known that scientific domains strongly differ in terms of citation rates, bibliometric indicators currently used in research assessment are often based on the sole use of raw citation numbers. This necessarily leads to unfair evaluation procedures, especially in cross-disciplinary contexts. For this reason, there is an increasing trend towards the formulation of normalization procedures able to suppress disproportions in citation numbers among scientific domains, and thus to lead to more fair cross-disciplinary evaluation criteria. In this paper, we rigorously test the performance of several field normalization procedures devoted to this purpose. We find that four procedures discussed in the literature do worse than the usual normalization with field averages. The latter drastically reduces citation disproportions among scientific disciplines. Finally, we find that a recently introduced two-parameters normalization scheme reduces citation disproportions to a level very close to the best achievable level of reduction.

Conference Topic
Scientometrics Indicators (Topic 1)

Introduction
The number of citations that a scientific paper has accumulated is often interpreted as a proxy of the influence of the same paper within the scientific community. Although the relation between citations and effective scientific influence is still under active debate (MacRoberts and MacRoberts, 1989,

MacRoberts and MacRoberts, 1986, Adler et al., 2009), and although the significance of citations is content- and discipline-dependent (Bornmann and Daniel 2008), citation numbers are often used in assessment exercises and their practical role in modern science is becoming more and more central. In the course of the last years, many bibliometric indicators have been developed with the aim of assessing the relevance of scientific research activities at different levels: journals (Garfield, 2006), scientists (Hirsch, 2005; Egghe, 2006), departments (Davis and Papanek, 2004), institutions (Kinney, 2007), etc. These indicators, however, are generally based on raw citation numbers, and thus have several limitations when used to perform comparisons across different fields of research. The limitation of the use of raw citation numbers appears evident already when used comparing two scientific papers. A paper in biochemistry typically accumulates more citations than a paper in mathematics but this does not necessarily imply that the former paper is more influential than the latter. Different scientific disciplines strongly differ in citation practices, and as a consequence the typical number of citations that a paper in a given field receives may strongly differ from the number of citations typical of another field.

To overcome this inherent disproportion in citation numbers among scientific fields, several approaches have been proposed to normalize citation numbers at the level of the single publication. The proposed schemes can be distinguished in two conceptually different classes:

(1) Target-based normalization: citation weights are functions of the cited papers. This class includes many different types of normalization techniques such as:

(i) Field averages (see inter alia Moed et al., 1985, 1988, 1995, Braun et al., 1985, Schubert et al., 1983, 1987, 1988, Schubert and Braun, 1986, 1996, and Vinkler 1986, 2003; see also Radicchi et al., 2008).
(ii) Average-based scalar difference from the mean (Glänzel, 2011).
(iii) Two-parameters reverse engineering (Radicchi and Castellano, 2012a).
(iv) Exchange rates (Crespo et al., 2012a, 2012b).

(2) Source-based normalization: citation weights are functions of the citing papers. The main example of this normalization scheme is represented by the so-called “fractional citation counting”, extensively studied by, inter alia, Zitt and Small, 2008, Moed, 2010, and Leydesdorff and Opthof, 2010, Glänzel et al., 2011, and Waltman et al., 2012.

While the development of cross-disciplinary citation indicators dates back to the 1980s, only recently scholars have started to apply them to large sets of empirical data, and statistically test their performances. Three methods have been proposed to quantitatively assess the performance of a generic normalization procedure:

(i) Between-group variance (Leydesdorff and Bormann, 2011).
(ii) Fairness test based on ranking (Radicchi and Castellano, 2012a).

(iii) Inequality due to Differences in Citation Practices method (IDCP) (Crespo et al., 2012a, 2012b).

Between-group variance is the simplest of the three tests, but, by construction, it vanishes for indicators normalized by field averages. This makes its applicability very limited. Although based on different principles, both the "fairness" and the IDCP tests leverage on strict statistical formalisms that do not require any strong assumption (i.e., they are distribution free statistical tests). The "fairness" test has already been applied to test the performance of indicators based on the two-parameters reverse engineering (Radicchi and Castellano, 2012b), field averages and fractional citation counts (Radicchi and Castellano, 2012a). It has been used also for testing the performances of normalized Impact Factors of journals. (Leydesdorff et al, 2012). The IDCP method has been used for field averages, exchange rates, and Glänzel type normalizations (Crespo et al, 2012a, 2012b).

In this paper, we perform an extensive analysis of several normalized indicators and assess their performance using the IDCP method. The dataset consists of publications appeared in different years spanning an interval of more than two decades and this allows also to analyse temporal trends in citation practices. Our results are in line with those already obtained in Radicchi and Castellano, 2012a and Radicchi and Castellano, 2012b according to which the reverse engineering procedure (Radicchi and Castellano, 2012b) outperforms other normalization methods.

Data

For our analysis, we make use of the same dataset as the one already analyzed in Radicchi and Castellano 2012b. This dataset is composed of six subsets, each including all publications appeared in 8,304 scientific journals in a distinct year of publication: 1980,1985,1990,1995, 1999 and 2004. Journal titles have been collected from the Journal of Citation Reports (JCR) database (http://science.thomsonreuters.com/cgi-bin/jrnlst/jlsubcatg.cgi?PC=D). For all publications we retrieved, from the Web Of Science (WOS, isiknowledge.com) database, in the period May 23-31, 2011, the number of citations they have accumulated (field "times cited"). We restrict our attention only to documents written in "English", and classified as "Article", "Letter", "Note" or "Proceedings Paper" for a total of 2,906,615 publications. Notice that the citation windows during which the papers in the different subsets have accrued citations are different, ranging from 31 years for the 1980 subset to 7 years for the papers published in 2004.

We further use the JCR classification of scientific journals in order to divide articles in different classes. JCR classification is composed of 172 subject-categories (also denoted as sub-fields in the following), each of them roughly representing a different research domain. As already emphasized by the same inventors (Pudovkin and Garfield 2002), the JCR classification is known to have

several weak points. One of them is that publications in the periodical literature are assigned to sub-fields via the journal in which they have been published. Many journals are assigned to a single sub-field, but many others are assigned to two, three, or even more sub-fields. For example in the datasets used in this paper, less than 2/3 of all articles are assigned to a single sub-field (to be more specific the percentage of articles assigned to a single subject-category varies with time: in 1980, the percentage of single-category papers was 67%, while only 56% in 2004).

To tackle this problem, two paths can be followed. The first is a *fractional* strategy, according to which each publication is fractioned into as many equal pieces as necessary, with each piece assigned to its corresponding sub-field. The second follows a *multiplicative* strategy in which each paper is counted as many times as necessary in the several sub-fields which it is assigned to. In this paper we adopt the multiplicative approach. This leads to a substantial increase in the total number of "papers": 42% in 1980, 45% in 1985, 48% in 1990, 56% in 1995, 58% in 1999 and 61% in 2004. However, we expect that our results will be comparable with those obtained with a fractional strategy, as demonstrated by Crespo et al. (2012b) for a dataset similar to the one used here.

Descriptive statistics

Descriptive statistics (available upon request) indicate that our yearly data present important differences in two respects: the distribution of documents by sub-field and the ratio of sub-field citation means to the overall citation mean do vary over the six years we study. However, within each year important similarities across sub-fields should be emphasized. For this purpose, we use the Characteristic Scores and Scales (CSS hereafter) technique, introduced by Schubert *et al.* (1987) in the analysis of citation distributions. Being scale- and size- invariant, the CSS method allows us to focus on the shape of sub-field citation distributions within each year.

The following *characteristic scores* are determined: μ_1 = mean citation for the entire yearly distribution, and μ_2 = mean citation for articles with citations above μ_1. Consider the partition of the distribution into three broad classes: articles with none or few citations below μ_1; fairly cited articles, with citations above μ_1 and below μ_2; and articles with a remarkable or outstanding number of citations above μ_2. In every year, we have computed the average and standard deviation over the 172 sub-fields for the percentage of articles in the three classes, as well as the corresponding statistics for the percentages of the total number of citations accounted by each class in every year. Since the results smoothly evolve during the 1980-2004 period, it suffices to report results in Table 1 for the two polar cases.

The small standard deviations in Table 1 indicate that sub-field citation distributions within each year share some fundamental characteristics: they are both similar and highly skewed in the sense that a large proportion of articles gets none or few citations while a small percentage of them account for a disproportionate amount of all citations. Specifically, for the two years in question between 69% and 73% of all articles receive citations below the mean and only account for, approximately, between 22% and 24% of all citations, while articles with a remarkable or outstanding number of citations represent about 8% or 10% of the total, and account for, approximately, between 42% and 47% of all citations. Other intermediate years present percentages comprised between those for 1980 and 2004. Finally, as can be seen in Table 1, these results closely resemble those concerning the shapes of citation distributions across a wide array of 219 sub-fields with a five-year citation window studied in Albarrán *et al.* (2011). As has been recently emphasized, this striking similarity between citation distributions paves the way for meaningful comparisons of citation counts across heterogeneous scientific disciplines (Radicchi et al. 2008, 2012a, 2012b, and Crespo et al. 2012a, 2012b).

Table 1. The skewness of science. Averages (and standard deviations) over 172 sub-field citation distributions in 1980 and 2004 *versus* previous results for articles published in 1998-2002 with a five-year citation window classified in 219 sub-fields

	Percentage of Articles In Category			Percentage of Total Citations Accounted For By Category		
	1	2	3	1	2	3
Results from our dataset, selected years:						
1. 1980	73.2 (4.3)	19.0 (2.6)	7.7 (2.1)	21.1 (5.0)	32.1 (2.3)	46.9 (5.5)
2. 2004	68.6 (3.5)	21.7 (2.0)	9.7 (1.7)	24.3 (3.6)	33.4 (1.4)	42.3 (3.5)
3. Previous results over 219 sub-fields for articles published in 1998-2002 with a five-year citation window. See Table 1, p. 391, in Albarrán *et al.* (2011, Table 1, p. 391):						
	68.6 (3.7)	-	10.0 (1.7)	29.1 (1.6)	-	44.9 (4.6)

Methods

Crespo et al. (2012a) introduced a simple model in which the number of citations received by an article is a function of two variables: the article's underlying scientific influence, and the field it belongs to. Consequently, the broad distribution of citation numbers for all articles in all fields –the all-sciences case– is the result of two components: differences in scientific influence within homogeneous fields, and differences in citation practices across fields.

In the implementation of this model using an additively decomposable inequality index, the citation inequality attributed to differences in citation practices is captured by a between-group inequality term in a certain partition by field and

citation quantile. We refer to (Crespo et al., 2012a, 2012b) for details. In practice one uses a citation inequality index *I*

$$I(C) = \frac{1}{N}\sum_{l}(c_l/\mu)\log(c_l/\mu)$$

where N is the total number of publications in a yearly dataset, c_l is the number of citations received by the l-th paper (l = 1,…, N), and μ is the mean of the distribution of c values. This quantity can be shown to be decomposed into the sum of three terms, one of them being the IDCP (Inequality due to Different Citation Practices)

$$ICDP = \sum_{\pi} \nu^{\pi} I(\mu^{\pi,1},\mu^{\pi,2},\ldots,\mu^{\pi,F})$$

where π denotes the quantile in the distribution for sub-field f = 1,…, F, $\nu^{\pi,f}$ is the share of total citations in quantile π of sub-field f, and $\nu^{\pi} = \Sigma_f \nu^{\pi,f}$.

The IDCP term captures the citation inequality attributable to differences in citation practices in different sub-fields. Thus, independently of the characteristics of citation distributions, the impact of any normalization procedure can be evaluated by the reduction in the IDCP term before and after normalization. The results of the method are dependent on the number of quantiles used to divide each citation distribution. In this paper, we use 100 quantiles (i.e., percentiles) in all our analysis. It is important to stress, however, that the quantitative difference due to the choice of the number of quantiles affects only the absolute values of the ICDP terms and not the comparison of the ICDP terms of different indicators and thus the measured level of reduction.

IDCP bounds

Raw citations – Raw citations are the basic information of impact that we have at our disposal. The IDCP term calculated on raw citation numbers quantifies the inequality of citation distributions due to different citation practices among scientific fields, and thus represents the term of reference for the computation of the reduction of such disproportion when using a normalized citation indicator. In this sense, the IDCP term calculated for raw citations represents the upper bound of the IDCP. Any reasonable normalized citation indicator must measure values of the IDCP below this upper bound.

Perfect normalization – What is the lowest bound for IDCP? In the case of infinite, real valued and identically distributed data, the ICDP term would be equal to zero. However, real citation numbers do not satisfy the former requirements and thus the best achievable ICDP term is in general larger than zero. When using ICDP to test the performance of normalized indicators, it is

therefore useful to compute the lowest value of ICDP that is achievable given the data, and use this value as a term of comparison for assessing the ability of the indicators to effectively remove differences between scientific sub-fields. In order to reach this goal, we use a very simple procedure. Given a sub-field, we assign to each paper with c citations a score s_c equal to the fraction of papers, within the same sub-field, that have accumulated a number of citations lower or equal to c. According to this rule scores have values in the range 0 to 1; scores preserve the natural order (including ties) of the original citation sequence; in each sub-field scores have exactly the same distribution, the uniform one. For these reasons, our way of assigning scores to papers represents a sort of "perfect normalization" scheme, and the ICDP term measured with these scores represents the best performance that can be achieved for a given data set.

Normalization procedures

Normalization by field average – We assign to each paper a score equal to $c_f = c/\mu_I$, where c is the number of citations accumulated by the paper and μ_I is the average number of citations received by papers in the same year of publication and in the same subject-category that was introduced in the section on descriptive statistics. Thus, c_f represents the relative impact, in terms of citations, of the paper within its field.

We additionally consider a slight variation of the former normalization scheme, where μ_I is calculated excluding uncited publications. This different approach has been used by Radicchi et al. 2008 and later suggested also by Waltman et al. 2011 and by Abramo et al. 2012 because it is supposed to lead to higher levels of reduction of citation disproportions among fields of science.

Normalization by median value – This represents a simple modification of the previous indicator, where the only difference is that the number of citations c received by a paper is divided by the median value μ_m of its field (instead of that by the average value μ_I). Since for some categories $\mu_m=0$, we calculate here the median citation number of each category by excluding uncited publications.

Normalization by two-parameters reverse engineering – Radicchi and Castellano 2012b have introduced a normalization scheme based on the use of two parameters. These parameters are estimated empirically on data, as the best estimates of the prefactor a and the exponent α of a power-law transformation able to make different citation distributions collapse on top of each other. This means that if the score of a paper is computed as $c' = (c/a)^{1/\alpha}$, with a and α parameters of the subject-category which the paper belongs to, then the distribution of c' is universal and no longer dependent on the specific subject-category considered. In particular, when two distributions have the same exponent, the transformation necessary for their collapse is linear, and the method

reduces to normalization by field average. Radicchi and Castellano, 2012b demonstrated that, for the vast majority of the sub-fields, the values of a and α are very similar, and the citation distributions are the same when plotted as a function of c'. A limited number of sub-fields, instead, is characterized by widely changing values of the transformation parameters, and thus have nonuniversal shapes of the distribution of c'.

Glanzel's normalization – This normalization involves the transformation of the raw data of any citation distribution with *N* papers, $\boldsymbol{c} = (c_1,\ldots, c_N)$, by the formula $c_i^* = c_i/(\mu_2 - \mu_1)$, where μ_1 and μ_2 are the two characteristic scores defined in the Descriptive statistics section.

Exchange rates normalization – Crespo et al. (2011a, b) find that the similarity of citation distributions allows the effect of idiosyncratic citation practices to be rather well estimated over a wide range of intermediate quantiles where citation distributions seem to differ by a scale factor. Consequently, a set of average-based measures, called exchange rates can profitably be estimated over that interval. Of course, this interval and the corresponding set of exchange rates must be estimated for each yearly sample.

Results

We apply the IDCP method to the indicators described above, calculated for the different publication years present in our dataset. We first analyze how the total citation inequality I and the IDCP term depend on time for raw citations.

Table 1. The evolution of total citation inequality and the IDCP term

	(1) Total citation Inequality	(2) IDCP	(3) = (2)/(1), in %
1980	1.058	0.124	11.7
1985	1.088	0.143	13.1
1990	1.030	0.139	13.5
1995	0.966	0.137	14.2
1999	0.890	0.120	13.4
2004	0.790	0.099	12.5

Table 1 shows that the absolute values of IDCP and I both tend to decrease over time, while their ratio remains approximately constant. This means that, in spite of the differences between the six yearly datasets, the relative importance of the differences in citation practices across sub-fields is of a similar order of magnitude, representing about 13% of total citation inequality. For articles published in 1998-2003 in 219 sub-fields with a five-year citation window, Crespo et al. (2011b) find that this percentage is 18%. For 22 broad fields, whose connection with the 219 or 172 Web of Science categories is unknown, this is approximately 14%.

In Figure 1, we plot the absolute value of IDCP when the different normalization procedures are applied, while in Table 2 we include the percentage that the IDCP term represents relative to the total citation inequality after normalization by the different procedures, excluding the perfect normalization that, understandably, does not perform well in relative terms because the total citation inequality (see Figure 1) is very low indeed.

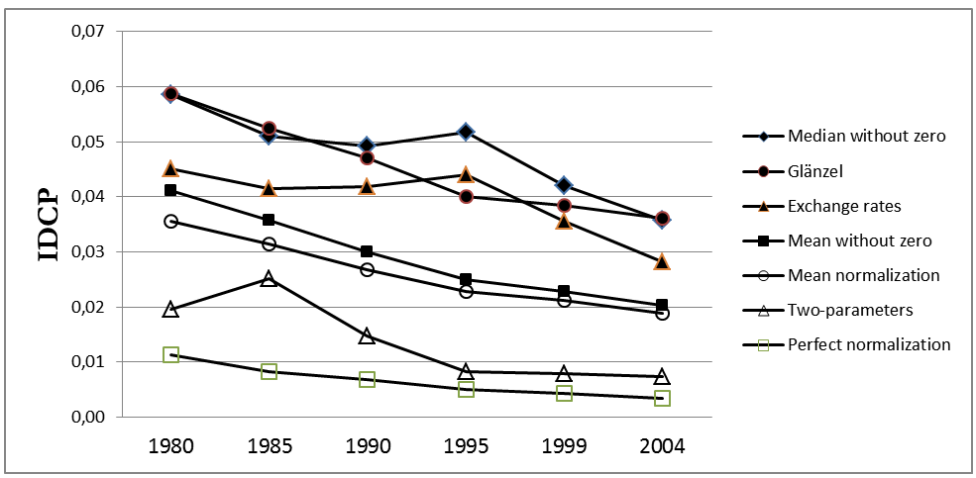


Figure 1. A comparison of the IDCP term in absolute value after applying the different normalization procedures

Table 2. A comparison of the percentage that the IDCP term represents relative to total citation inequality after applying the different normalization procedures

	Median Without 0s	Glänzel	Exchange Rates	Mean Without 0s	Mean	Two Parameters
1980	5.6	6.2	4.4	4.3	3.6	1.7
1985	4.9	5.5	4.0	3.7	3.2	2.1
1990	5.0	5.3	4.3	3.3	2.9	1.3
1995	5.6	4.8	4.8	3.0	2.7	0.8
1999	4.9	4.9	4.2	2.9	2.7	0.8
2004	4.8	5.2	3.8	2.9	2.7	0.9

The contest offers clear results. Normalization by sub-field mean citations dominates four other alternatives. However, the two-parameters scheme is the one that gets closer to the perfect normalization benchmark.

Further insight into the origin of the variation of the performance of the best normalization procedures is provided by Figure 2, where the quantity $I(\pi)$, capturing the citation inequality due to differences in citation practices across sub-

fields in every percentile, is plotted as a function of the percentile π for papers published in 1990 (similar results occurring for other publication years) Because I(π) is too large for many low percentiles and some very high ones, Figure 2 only reports results for the interval (50, 96).

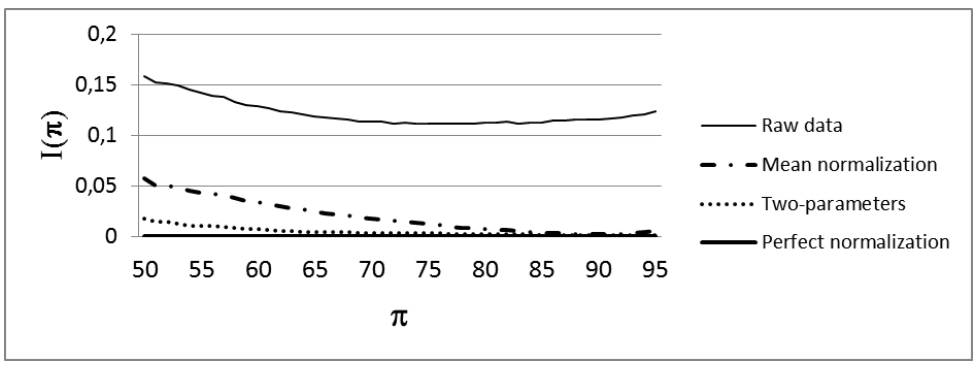


Figure 2. The citation inequality due to differences in citation practices across sub-fields in every percentile, I(π), as a function of π□□□□□□□□□□R□□□□□□□□□ and after the best normalization procedures are applied for the 1990 dataset.

The following comments are in order. Firstly, as in Crespo et al. (2012a, b), the quantity I(π) for the raw data is relatively constant over a large quantile interval (although only for high values of π). This is what allows us to define a set of average-based exchange rates over that interval. Secondly, the reduction of I(π) achieved by the best normalization procedures is well illustrated.

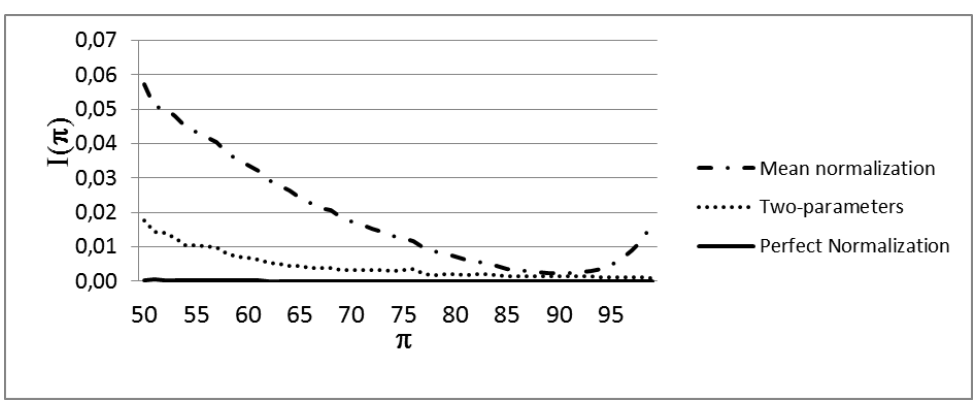


Figure 3. The comparison of the effect on I(π) caused by the best normalization procedures for the 1990 dataset

Finally, Figure 3 amplifies Figure 2 in order to appreciate the differences between the best two normalization alternatives. At the very upper tail of citation distributions –namely, when it most matters– the mean normalization's

performance clearly worsens. This is also the case with the alternative procedures not included in Figure 4. However, the two-parameters scheme does extremely well in that interval.

Discussion and Conclusions

While the use of citation numbers in research assessment exercises is becoming more and more relevant, there is still much room for the improvement of bibliometric indicators devoted to the quantification of research impact. In particular, there is a strong necessity to find proper ways of suppressing disproportions in raw bibliometric measures merely due to different citation practices in different fields. In this paper we have presented a quantitative assessment of the effectiveness of several procedures for normalizing raw citation numbers.

Using the recently introduced IDCP method, we have measured the performance of the different procedures applied to papers published in different years ranging from 1980 to 2004 divided in 172 distinct sub-fields. It turns out that:

- For raw citation numbers the total inequality of citation distributions and the inequality to due different citation practices both tend to decrease over time, while their ratio remains approximately constant.
- Among the different normalization procedures, the recently introduced reverse engineering transformation based on two-parameters performs better than the others, but also the normalization by field averages yields good results.
- The two-parameter procedure outperforms other methods (and is close to the perfect normalization benchmark) in particular at the upper tail of the citation distributions, i.e. for highly cited publications.

These results clearly indicate that the regularity of the features of citation distributions described in Albarrán et al. (2011a, 2011b), as well as in the Descriptive statistics section of this paper, can be fruitfully used for the formulation of normalization procedures that are able to drastically reduce the disproportions in raw citation counts among different sub-fields of science. At the same time, however, much work is still needed for this basic but central problem. Apart from further study of the robustness of the available results, possibly the more important issue is the development of better classification schemes able to define fields and sub-field of science in a more coherent manner.

References

Abramo, G., Cicero, T. & D'Angelo, C.A. (2012) How important is choice of the scaling factor in standardizing citations? *Journal of Informetrics*, 6, 645–654.

Adler, R. Ewing, J. & Taylor P. (2009). Citation statistics. *Statistical Science,* 24, 1–14.

Albarrán, P., & Ruiz-Castillo, J. (2011a) References Made and Citations Received By Scientific Articles. *Journal of the American Society for Information Science and Technology*, 62, 40-49.

Albarrán, P., Crespo, J., Ortuño, I., & Ruiz-Castillo, J. (2011b) The Skewness of Science In 219 Sub-fields and a Number of Aggregates. *Scientometrics*, 88, 385-397.

Bornmann L. & Daniel H.D. (2008). What do citation counts measure? A review of studies on citing behavior. *Journal of Documentation*, 64, 45–80.

Braun, T., W. Glänzel, & A. Schubert (1985), Scientometrics Indicators. A 32 Country Comparison of Publication Productivity and Citation Impact. World Scientific Publishing Co. Pte. Ltd., Singapore, Philadelphia.

Crespo, J. A., Li, Yunrong, & Ruiz-Castillo, J. (2012a) Differences in Citation Impact Across Scientific Fields. Working Paper 12-06, Universidad Carlos III (http://hdl.handle.net/10016/14771).

Crespo, J. A., Li, Yunrong, Herranz, N., & Ruiz-Castillo, J. (2012b) Field Normalization at Different Aggregation Levels, Working Paper 12-022, Universidad Carlos III (htpp:/hdl.handle.net/10016/15344).

Davis, P. & Papanek, G.F. (1984). Faculty ratings of major economics departments by citations. *American Economical Review*, 74, 225–230.

Egghe, L. (2006). Theory and practise of the g-index. *Scientometrics*, 69, 131–152.

Garfield, E. (2006). The history and meaning of the journal impact factor. *Journal of the American Medical Association*, 295, 90–93.

Glänzel, W. (2011) The Application of Characteristic Scores and Scales to the Evaluation and ranking of Scientific Journals. *Journal of Information Science*, 37, 40-48.

Glänzel, W., Schubert, A., Thijs, B., & Debackere, K. (2011) A priori vs. a posteriori normalization of citation indicators. The case of journal ranking. *Scientometrics*, 87, 415-424.

Hirsch, J.E. (2005). An index to quantify an individual's scientific research output. *Proceedings of the National Academy of Science USA*, 102, 16569–16572.

Kinney, A.L. (2007). National scientific facilities and their science impact on nonbiomedical research. *Proceedings of the National Academy of Science USA*, 104, 17943–17947.

Leydesdorff, L. & Opthof, T. (2010) Normalization at the Field level: Fractional Counting of Citations, *Journal of Informetrics*, 4, 644-646.

Leydesdorff, L., Radicchi, F., Bornmann, L., Castellano, C., & de Nooye, W. (2012) Field-normalized Impact Factors: A Comparison of Rescaling versus Fractionally Counted Ifs. in press, *Journal of the American Society for Information Science and Technology*.

MacRoberts, M.H. & MacRoberts, B.R. (1989). Problems of citation analysis: Acritical review. *Journal of the American Society for Information Science* and Technology, 40, 342–349.

MacRoberts, M.H. & MacRoberts, B.R. (1996). Problems of citation analysis. *Scientometrics* 36: 435–444.
Moed, H. F., Burger, W.J. Frankfort, J.G., & van Raan, A.F.J. (1985) The Use of Bibliometric Data for the Measurement of University Research Performance. *Research Policy*, 14, 131-149.
Moed, H. F., & van Raan, AF.J. (1988) Indicators of Research Performance. in A. F. J. van Raan (ed.), Handbook of Quantitative Studies of Science and Technology, North Holland: 177-192.
Moed, H. F., De Bruin, R.E, & van Leeuwen, Th.N. (1995) New Bibliometrics Tools for the Assessment of national Research Performance: Database Description, Overview of Indicators, and First Applications. *Scientometrics*, 33, 381-422.
Pudovkin, A.I. & Garfield, E. (2002). Algorithmic procedure for finding semantically related journals. *Journal of the American Society for Information Science* and Technology, 53, 1113–1119.
Radicchi, F., Fortunato, S., & Castellano, C. (2008) Universality of Citation Distributions: Toward An Objective Measure of Scientific Impact. *Proceedings of the National Academy of Science USA*, 105, 17268-17272.
Radicchi, F., & Castellano, C. (2012a) Testing the fairness of citation indicators for comparisons across scientific domains: The case of fractional citation counts. *Journal of Informetrics*, 6, 121-130.
Radicchi, F., & Castellano, C. (2012b) A Reverse Engineering Approach to the Suppression of Citation Biases Reveals Universal Properties of Citation Distributions. *Plos One*, 7, e33833.
Schubert, A., & Braun, C. (1986) Relative Indicators and Relational Charts for Comparative Assessment of Publication Output and Citation Impact. *Scientometrics*, 9, 281-291.
Schubert, A., & Braun, T. (1996) Cross-field Normalization of Scientometric Indicators. *Scientometrics*, 36, 311-324.
Schubert, A., Glänzel, W., Braun, T. (1983) Relative Citation Rate: A New Indicator for Measuring the Impact of Publications. in D. Tomov and L. Dimitrova (eds.), Proceedings of the First National Conference with International Participation in Scientometrics and Linguistics of Scientific Text, Varna.
Schubert, A., Glänzel, W., & Braun, T. (1987) A New Methodology for Ranking Scientific Institutions. *Scientometrics*, 12, 267-292.
Schubert, A., Glänzel, W., & Braun, T. (1988) Against Absolute Methods: Relative Scientometric Indicators and Relational Charts as Evaluation Tools. in A. F. J. van Raan (ed.), Handbook of Quantitative Studies of Science and Technology: 137-176.
Vinkler, P. (1986) Evaluation of Some Methods For the Relative Assessment of Scientific Publications. *Scientometrics*, 10, 157-177.
Vinkler, P. (2003) Relations of Relative Scientometric Indicators. *Scientometrics*, 58, 687-694.

Waltman, L, van Eck, N. J., & Van Raan, A.F.J. (2011) Universality of citation distributions revisited. *Journal of the American Society for Information Science and Technology*, 63, 72-77.

Zitt M., & Small H. (2008) Modifying the journal impact factor by fractional citation weighting: The audience factor. *Journal of the American Society for Information Science and Technology*, 59, 1856-1860.

A RELATION BETWEEN POWER LAW DISTRIBUTIONS AND HEAPS' LAW

Shi Shan[1], DingHua Shi[2] and YiFei Zhang[3]

[1] *shanshihill@126.com*
Shanghai University, Center for Information Studies, 99 Shangda Road, 200444 Shanghai (China)

[2] *shidh2001@263.net*
Shanghai University, School of Science, 99 Shangda Road, 200444 Shanghai (China)

[2] *zhangyifei@shu.edu.cn*
Shanghai University, School of Management, 99 Shangda Road, 200444 Shanghai (China)

Abstract
The growth of vocabulary size as a concavely increasing function of the text length (its number of words) has been described in Heaps' law. Furthermore, word frequencies in texts can be reasonably described by power law distributions. A particular interesting phenomenon is the coexistence of Heaps' law and power law distributions. In this paper we prove that (under a weak condition) Heaps' law is equivalent to the power law distribution with exponent 1-θ θ∈(0,1).

Conference Topic
Bibliometrics in Library and Information Science (Topic 14) and Webometrics (Topic 7).

Introduction
How do actual texts evolve with text length? What are "normal" growth patterns in languages and information networks? Two models describing real text generation are worth mentioning. The first one, a formulation put forward by Zipf (Zipf, 1936) and is now widely known as Zipf's law, establishes that the number of words $f(k)$ that occur exactly k times in a text decays with k as $f(k) = ck^{-(1+\alpha)}$, where $c > 0$ and $\alpha > 0$. This power law size-frequency relation indicates a power law probability distribution of the size k itself, say $p(k) = ck^{-(1+\alpha)}$. The second model is due to Heaps (Heaps, 1978), who showed that vocabulary size $n(k)$ grows in a concave function with text size k, namely $n(k) = ck^{1-\theta}$ with $\theta \in (0,1)$ and $c > 0$. A particular interesting phenomenon is the coexistence of the power law distribution and Heaps' law. Besides the statistical regularities of text, words contained by web pages resulted from web searching (Lansey and Bukiet, 2009) and keywords for scientific publications also

simultaneously display the power law distribution and Heaps' law. In particular, the power law distribution and Heaps' law are closely related to the evolving networks. It is well known that some networks grow in an accelerating manner (Lu, Zhang & Zhou, 2010) and have power law structures, in fact, the former property corresponds to Heaps' law that the number of nodes grows in a concavely increasing function with the total degree of nodes, while the latter is equivalent to the power law distribution.

Based on a variant of Simon model (Simon, 1955), Zanette and Montemurro (Zanette and Montemurro, 2005) showed that Mandelbrot's law is a result from Heaps' law. By using a more polished approach, Leijenhorst and Weide (Leijenhorst and Weide, 2005) provided a formal derivation of Heaps' law from Mandelbrot's law. Using an exact informetric argument on random sampling in the items, Egghe (Egghe, 2007) showed that, in most cases, $n(k)$ is a concavely increasing function, in accordance with practical examples. Nevertheless, it appears that the story of Heaps' Law is inseparably connected with its scientometric application from the very beginnings up to the recent days. Zhang, Lü, Liu and Zhou (Zhang et al, 2008) found that there is a power law correlation between the cumulative number of distinct keywords and the cumulative number of keyword occurrences. They also monitored the decay trend of most popular keywords. Interestingly, top journals from various subjects share very similar decaying tendency, while journals with low impact indexes exhibit completely different behaviour (Zhang et al, 2008). In this paper, we introduce the notion of gradual variation for function on N_+, which plays a key role in studying power law distributions. Some necessary and sufficient conditions for power law distribution are obtained. In particular, we show an example that $G(k)=\sum_{i\geq k} p(i)=\eta(k)k^{-\alpha}\sim d>0$, but it does not hold that $p(k)\sim ck^{-(\alpha+1)}$, where $c>0$ and $\alpha>0$. Finally, we prove that (under a weak condition) Heaps' law is equivalent to the power law distribution with exponent $1-\theta$ ($\theta\in(0,1)$).

The Power Law Distribution

We start with definitions and properties of the power law distribution and the gradually varying function, the latter plays a key role in our discussion.

Here and throughout the paper, $p(k)$ is a positive probability on $N_+=\{1,2,\cdots\}$. $F(\cdot)$ is a distribution function, i.e., $F(k)=\sum_{i\leq k} p(i)$. $G(k)$ is a complementary distribution function, (or right cumulative distribution function), i.e., $G(k)=\sum_{i\geq k} p(i)$. $\log(\cdot)$ stands for the natural logarithm on the set of positive reals, i.e., $\log k=\int_1^k \frac{1}{x}dx$. $\Gamma(\alpha)$ denotes the usual gamma function, i.e.,

$\Gamma(\alpha)=\int_0^\infty x^{\alpha-1}e^{-x}dx$ for $\alpha>0$. The relation of asymptotic equivalence between $f(k)$ and $g(k)$, written by $f(k)\sim g(k)$, means that the ratio of $f(k)$ to $g(k)$ goes to 1 as $k\to\infty$. To write that $f(k)=o(g(k))$ means that the ratio of $f(k)$ to $g(k)$ goes to 0 as $k\to\infty$.

We shall need the following two properties and an auxiliary result.

Property 2.1 $(1+\frac{1}{k})^\alpha-1=\frac{\alpha}{k}+o(\frac{1}{k})$ for $k\in N_+$. (2.1)

Property 2.2 If $\xi(k)=o(1)$, $\log(1+\xi(k))\sim\xi(k)$. (2.2)

Stolz Theorem Let $g(k)$ and $f(k)$ be functions on N_+. (1) If (i) $f(k)=o(1)$; (ii) $g(k)=o(1)$ and there exists $m\in N_+$ such that $g(k+1)<g(k)$ for all $k>m$; and (iii) the limit $\lim_{k\to\infty}\frac{f(k+1)-f(k)}{g(k+1)-g(k)}$ exists; it holds that $\lim_{k\to\infty}\frac{f(k)}{g(k)}=\lim_{k\to\infty}\frac{f(k+1)-f(k)}{g(k+1)-g(k)}$. (2) If (i) $g(k)\to\infty$ (as $k\to\infty$); (ii) there exists $m\in N_+$ such that $g(k+1)>g(k)$ for all $k>m$; and (iii) the limit $\lim_{k\to\infty}\frac{f(k+1)-f(k)}{g(k+1)-g(k)}$ exists; it holds that $\lim_{k\to\infty}\frac{f(k)}{g(k)}=\lim_{k\to\infty}\frac{f(k+1)-f(k)}{g(k+1)-g(k)}$.

Refer for the above to (Klambauer, 1975).

Definition 2.1 A positive probability $p(k)$ on N_+ is the power law distribution with exponent α, if $p(k)=\psi(k)k^{-(\alpha+1)}$, where $\psi(k)\to c>0$ (as $k\to\infty$) and $\alpha>0$.

Definition 2.2 A positive function $\eta(k)$ on N_+ varies gradually (at ∞) if and only if $\eta(k+1)/\eta(k)=1+o(1/k)$.

Theorem 2.1 $p(k)$ is the power law distribution with exponent α if and only if $G(k)=\eta(k)k^{-\alpha}$, where $G(k)=\sum_{i\geq k}p(i)$, $\eta(k)$ varies gradually and $\eta(k)\to d>0$ (as $k\to\infty$), and $\alpha>0$.

Proof. By $G(k)-G(k+1)=p(k)=\psi(k)k^{-(\alpha+1)}$, where $\psi(k)\to c$ (as $k\to\infty$), we have

$$\frac{G(k)-G(k+1)}{k^{-\alpha}-(k+1)^{-\alpha}}=\frac{c(1+o(1))k^{-1}}{1-(1-\frac{1}{k})^{\alpha}}.$$

Using (2.1), we get

$$\frac{G(k)-G(k+1)}{k^{-\alpha}-(k+1)^{-\alpha}}\sim\frac{c}{\alpha}. \tag{2.3}$$

Therefore, by Stolz Theorem,

$$\frac{G(k)}{k^{-\alpha}}\sim\frac{c}{\alpha}, \tag{2.4}$$

i.e., $G(k)=\eta(k)k^{-\alpha}$, where $\eta(k)\to d=c/\alpha$ (as $k\to\infty$).

Otherwise, from

$$\begin{aligned}\frac{G(k+1)}{G(k)}&=\frac{G(k)-p(k)}{G(k)}\\&=1-\frac{c(1+o(1))k^{-(\alpha+1)}}{d(1+o(1))k^{-\alpha}}\\&=1-\frac{\alpha}{k}+o(\frac{1}{k})\end{aligned} \tag{2.5}$$

and $\eta(k)=G(k)k^{-\alpha}$, we have

$$\begin{aligned}\frac{\eta(k+1)}{\eta(k)}-1&=(1+\frac{1}{k})^{\alpha}\frac{G(k+1)}{G(k)}-1\\&=(1+\frac{1}{k})^{\alpha}(1-\frac{\alpha}{k}+o(\frac{1}{k}))-1\\&=o(\frac{1}{k}).\end{aligned}$$

This shows that $\eta(k)$ is a gradually varying function.

Conversely, seeing (2.5) and $\frac{G(k)}{G(k+1)}-1=\frac{p(k)}{G(k)}$, we have

$$\frac{G(k)}{G(k+1)}-1=\frac{\alpha}{k}+o(\frac{1}{k}),$$

$$\begin{aligned}p(k)&=(\frac{\alpha}{k}+o(\frac{1}{k}))G(k+1)\\&=(\frac{\alpha}{k}+o(\frac{1}{k}))(1-\frac{\alpha}{k}+o(\frac{1}{k}))G(k)\end{aligned}$$

And $p(k)=\frac{\alpha}{k}(1+o(1))\eta(k)k^{-\alpha}$. Noting that $\eta(k)\to d=c/\alpha$ (as $k\to\infty$) (see(2.4)) and letting $\psi(k)$ be equal to $\alpha\eta(k)(1+o(1))$, we get

$$p(k)=\psi(k)k^{-(\alpha+1)},$$

with $\psi(k)\to c$ (as $k\to\infty$).

Corollary 2.1 Let $p(k)$ be a positive probability on N_+ and $G(k)=\sum_{i\geq k}p(i)$. Then $k^{1+\alpha}p(k)=\psi(k)\to c$ (as $k\to\infty$) if and only if $k^{\alpha}G(k)=\eta(k)\to d=c/\alpha$ (as $k\to\infty$) and $\eta(k)$ varies gradually, where $\alpha>0$ and $c>0$.

Proof. It is easy to check from the proof of Theorem 2.1.

Remark 2.1 Corollary 2.1 shows an interesting property that $k^{1+\alpha}p(k)=\psi(k)\to c$ and $k^{\alpha}G(k)=\eta(k)\to c/\alpha$ as $k\to\infty$. Comparing with the continuous power law distribution, the Pareto distribution on $[\varepsilon,\infty)$ ($\varepsilon>0$), we have

$$x^{\alpha+1}f(x)=c=\alpha\varepsilon^{\alpha}$$

and

$$x^{\alpha}G(x)=\frac{c}{\alpha}=\varepsilon^{\alpha},$$

where $f(x)$ is the density function and $G(x)=\int_x^{\infty}f(y)dy$.

Example 2.1 For the Waring distribution

$$p(k)=\alpha\frac{\Gamma(\alpha+\beta)\Gamma(\beta+k-1)}{\Gamma(\beta)\Gamma(\beta+k+\alpha)},\ (\alpha>0,\ \beta>0,\ k\in N_+)$$

It follows by induction that

$$G(k)=\frac{\Gamma(\alpha+\beta)\Gamma(\beta+k-1)}{\Gamma(\beta)\Gamma(\alpha+\beta+k-1)}.$$

Let $\eta(k)=G(k)k^{\alpha}$. From the fact that $\alpha\Gamma(\alpha)=\Gamma(\alpha+1)$ and then

$$\frac{\eta(k+1)}{\eta(k)}=\frac{G(k+1)}{G(k)}(1+\frac{1}{k})^{\alpha}$$

$$=(1-\frac{\alpha}{\alpha+\beta+k-1})(1+\frac{\alpha}{k}+o(\frac{1}{k})),$$

We get

$$\frac{\eta(k+1)}{\eta(k)}-1=\frac{\alpha}{k}+o(\frac{1}{k})-\frac{\alpha}{\alpha+\beta+k-1}(1+o(1)),$$

Which means that $\eta(k+1)/\eta(k)=1+o(1/k)$. Thus $\eta(k)$ is a gradually varying function. Using Stirling's formula: $\Gamma(\alpha+k)/\Gamma(k)\sim k^{\alpha}$ (as $k\to\infty$), we have that $\eta(k)\to\Gamma(\alpha+\beta)/\Gamma(\beta)$ (as $k\to\infty$) and $\eta(k)$ varies gradually, and $G(k)=\eta(k)k^{-\alpha}$.

Theorem 2.1 leads us to the following definition

Definition 2.3 Let $p(k)$ be a positive probability on N_+ and $G(k)=\sum_{i\geq k}p(i)$. $p(k)$ is the power law distribution with exponent α if $G(k)=\eta(k)k^{-\alpha}$, where $\eta(k)\to d>0$ (as $k\to\infty$) and $\eta(k)$ varies gradually, and $\alpha>0$.

Theorem 2.2 Let $p(k)$ be a positive probability on N_+ and $F(k)=\sum_{i\leq k}p(i)$. Then $F(k)$ varies gradually if and only if $p(k)=o(\frac{1}{k})$.

Proof. Since

$$k(\frac{F(k+1)}{F(k)}-1)=k\frac{p(k+1)}{F(k)}\sim(k+1)p(k+1),$$

It follows that $F(k+1)/F(k)=1+o(1/k)$ if and only if $p(k)=o(1/k)$.

Example 2.2 Let $p(k)$ be the geometric distribution on N_+, i.e., $p(k)=\theta(1-\theta)^{k-1}$, where $\theta\in(0,1)$. For $\lambda=-\log(1-\theta)>0$, it is easy to check that $kp(k)=\theta ke^{-\lambda(k-1)}=o(1)$ and then $F(k)=1-(1-\theta)^{k-1}$ is a gradually varying function.

Remark 2.2 As Example 2.2 shows, being a gradually varying function for $F(k)$ is a weak assumption without involving some strong conditions such as heavy tails, thick tails or power law style for distributions.

Theorem 2.3 $p(k)$ is the power law distribution with exponent α if and only if $\Pi(k)=\frac{kp(k)}{G(k)}\to\alpha$ (as $k\to\infty$) and $G(k)k^{\alpha}=\eta(k)\to d>0$ (as $k\to\infty$).

Proof. Now $G(k)k^{\alpha} = \eta(k) \to d > 0$ (as $k \to \infty$) and

$$\Pi(k) = \frac{kp(k)}{G(k)} = k(1 - \frac{G(k+1)}{G(k)}) \to \alpha \text{ (as } k \to \infty).$$

This shows that

$$\frac{G(k+1)}{G(k)} = 1 - \frac{\alpha}{k} + o(\frac{1}{k}).$$

From $G(k) = \eta(k)k^{-\alpha}$, it follows that

$$\begin{aligned}\frac{\eta(k+1)}{\eta(k)} - 1 &= (1+\frac{1}{k})^{\alpha}\frac{G(k+1)}{G(k)} - 1 \\ &= (1+\frac{1}{k})^{\alpha}(1 - \frac{\alpha}{k} + o(\frac{1}{k})) - 1 \\ &= o(\frac{1}{k})\end{aligned}$$

and $\eta(k)$ varies gradually (at ∞).

Conversely, from Corollary 2.1 the fact that

$$k^{\alpha+1}p(k) = \psi(k) \to c \text{ (as } k \to \infty)$$

implies $k^{\alpha}G(k) = \eta(k) \to d = c/\alpha$ (as $k \to \infty$). Then $\Pi(k) = \frac{\psi(k)}{\eta(k)} \to \alpha$ (as $k \to \infty$).

Proposition 2.1 (1) Both $\eta_1(k)$ and $\eta_2(k)$ are gradually varying functions, then $\eta(k) = \eta_1(k)\eta_2(k)$ varies gradually. (2) $\overline{\eta}(k)$ is the gradually varying function, then $\eta(k) = (\overline{\eta}(k))^{-1}$ varies gradually.

Proof. (1) From

$$\begin{aligned}\frac{\eta(k+1)}{\eta(k)} - 1 &= \frac{\eta_1(k+1)\eta_2(k+1)}{\eta_1(k)\eta_2(k)} - 1 \\ &= (1+o(\frac{1}{k}))(1+o(\frac{1}{k})) - 1 \\ &= o(\frac{1}{k}),\end{aligned}$$

$\eta(k)$ then varies gradually. (2) It is easy to check from the definition of gradual variation.

Extended Heaps' Law

The power law relation between the text length k and the number of different words in a text $n(k)$, $n(k) = \bar{c}k^{1-\theta}$ ($\bar{c} > 0$), is usually referred to as Heaps' Law. The sub-linear growth of k with $n(k)$ is insured if the Heaps exponent is more than zero and less than one. Let $e(k)$ be the expectation number of occurrences of a word in a text containing k words, we have that $e(k) = \frac{k}{n(k)} = ck^{\theta}$ with $c\bar{c} = 1$, if Heaps' law holds. Let X be a random variable, "$X = k$"($k \in N_+$) the event "a word occurs k times" and $E(X \mid X \le k)$ the expectation of X under the condition of the text length being k or a word occurring k times at most from the above, we have that $E(X \mid X \le k) = ck^{\theta}$, if Heaps' law holds. By replacing the constant $\bar{c}$ in the original form of Heaps' law with a gradually varying function $\bar{\varphi}(k)$ with $\lim_{k\to\infty} \bar{\varphi}(k) = \bar{c}$, $n(k) = \bar{\varphi}(k)k^{1-\theta}$ is an extension of Heaps' law, and the corresponding conditional expectation becomes $E(X \mid X \le k) = \varphi(k)k^{\theta}$, where $\varphi(k)\bar{\varphi}(k) = 1$, $\varphi(k)$ varies gradually (see Proposition 2.1) and $\varphi(k) \to c = \frac{1}{\bar{c}}$ (as $k \to \infty$).

Definition 3.1 Let $N(k) = \frac{k}{E(X \mid X \le k)}$. $N(k)$ is extended Heaps' law on N_+, if $N(k) = \bar{\varphi}(k)k^{1-\theta}$, where $\theta \in (0,1)$, $\bar{\varphi}(k)$ varies gradually and $\bar{\varphi}(k) \to \bar{c} > 0$ as $k \to \infty$.

Let $\lambda(k) = \sum_{i \le k} ip(i)$, then $\lambda(k) = F(k)E(X \mid X \le k)$, where $F(k) = \sum_{i \le k} p(i)$.

Lemma 3.1 Let $\lambda(k) = F(k)\varphi(k)k^{\theta}$ ($\theta \in (0,1)$), where $F(k)$ is the gradually varying function, $\varphi(k)$ is the gradually varying function and $\varphi(k) \to c$ as $k \to \infty$, then $G(k) = \eta(k)k^{-(1-\theta)}$, where $\eta(k)$ varies gradually and $\eta(k) \to \frac{c\theta}{1-\theta}$ as $k \to \infty$, i.e., $p(k)$ is the power law distribution with exponent $1-\theta$.

Proof. Now $F(k)\varphi(k)$ is the gradually function(from the assumption and proposition 2.1), i.e.,

$$\frac{F(k+1)\varphi(k+1)}{F(k)\varphi(k)} = 1 + o(\frac{1}{k}),$$

and

$$\lambda(k+1)-\lambda(k)= F(k+1)\varphi(k+1)(k+1)^{\theta}-F(k)\varphi(k)k^{\theta}$$
$$= F(k)\varphi(k)k^{\theta}(\frac{F(k+1)\varphi(k+1)}{F(k)\varphi(k)}(1+\frac{1}{k})^{\theta}-1),$$

and $\lambda(k+1)-\lambda(k)=(k+1)p(k+1)$, it follows that

$$(k+1)p(k+1)= F(k)\varphi(k)k^{\theta}((1+o(\frac{1}{k}))(1+\frac{\theta}{k}+o(\frac{1}{k}))-1)$$
$$= F(k)\varphi(k)k^{\theta}(\frac{\theta}{k}+o(\frac{1}{k}))$$
$$= F(k)\varphi(k)\theta k^{-(1-\theta)}(1+o(1))$$

and

$$p(k+1)=F(k)\varphi(k)\frac{\theta}{k+1}k^{-(1-\theta)}(1+o(1)). \tag{3.1}$$

Substituting $F(k+1)=F(k)(1+o(\frac{1}{k}))$ and $\varphi(k+1)=\varphi(k)(1+o(\frac{1}{k}))$ into (3.1) yields

$$p(k+1)=F(k+1)\varphi(k+1)\theta k^{-(2-\theta)}(1+o(1)),$$

namely,

$$p(k)=F(k)\varphi(k)\theta(1+o(1))k^{-(2-\theta)}.$$

Let $\psi(k)=F(k)\varphi(k)\theta(1+o(1))$, then $\psi(k)\to\theta c$ (as $k\to\infty$). $p(k)$ is the power law distribution with $p(k)=\psi(k)k^{-(2-\theta)}$. Using Theorem 2.2, we have $G(k)=\sum_{i\geq k}p(i)=\eta(k)k^{-(1-\theta)}$, where $\eta(k)$ varies gradually and $\eta(k)\to\frac{\theta c}{1-\theta}$ (as $k\to\infty$).

Lemma 3.2 Let $G(k)=\eta(k)k^{-(1-\theta)}$ ($\theta\in(0,1)$), where $\eta(k)\to\frac{\theta c}{1-\theta}$ (as $k\to\infty$), $\eta(k)$ varies gradually, then $\lambda(k)=F(k)\varphi(k)k^{\theta}$, where $F(k)$ is the gradually varying function, $\varphi(k)\to c$ (as $k\to\infty$) and $\varphi(k)$ is the gradually varying function.

Proof. From $\lambda(k+1)-\lambda(k)=(k+1)p(k+1)$, and Theorem 1.1 and Corollary 1.1, we have

$$\lambda(k+1)-\lambda(k)=\frac{\eta(k+1)(1-\theta)}{(k+1)^{1-\theta}}(1+o(1)).$$

Define $\zeta(k)=\frac{1}{\theta}\eta(k)k^{\theta}$, then

$$\begin{aligned}\zeta(k+1)-\zeta(k)&=\frac{1}{\theta}\eta(k)k^{\theta}((1+\frac{1}{k})^{\theta}\frac{\eta(k+1)}{\eta(k)}-1)\\&=\frac{1}{\theta}\eta(k)k^{\theta}(\frac{\theta}{k}+o(\frac{1}{k}))\\&=\eta(k)k^{-(1-\theta)}(1+o(1)).\end{aligned}$$

There exists $\overline{m}\in N_{+}$ such that for all $k\in N_{+}$ and $k>\overline{m}$, $\zeta(k+1)-\zeta(k)>0$, and $\zeta(k)\to\infty$ (as $k\to\infty$) (because $\eta(k)\to d$ as $k\to\infty$ and from Proposition 2.1). Now

$$\frac{\lambda(k+1)-\lambda(k)}{\zeta(k+1)-\zeta(k)}=(1+o(1))\frac{\eta(k+1)}{\eta(k)}(1-\frac{1}{k+1})^{1-\theta}(1-\theta)\to(1-\theta)\ (\text{as } k\to\infty),$$

from the Stolz theorem, it follows that

$$\frac{\lambda(k)}{\zeta(k)}=(1-\theta)(1+o(1))$$

or

$$\begin{aligned}\lambda(k)&=\zeta(k)(1-\theta)(1+o(1))\\&=\frac{1-\theta}{\theta}\eta(k)k^{\theta}(1+o(1)).\end{aligned}$$

Denoting $\varphi(k)$ by $\frac{1-\theta}{\theta}\eta(k)(1+o(1))$ and seeing the condition that $\eta(k)\sim\frac{c\theta}{1-\theta}$, we conclude that $\lambda(k)=\varphi(k)k^{\theta}$, where $\varphi(k)\to c$ (as $k\to\infty$). Next we will show that $\varphi(k)$ is the gradually varying function.

From $k^{1-\theta}G(k)=\eta(k)\to\frac{c\theta}{1-\theta}$ (as $k\to\infty$) and Corollary 2.1, we have $k^{2-\theta}p(k)=\psi(k)\to c\theta$ (as $k\to\infty$). Let $(k)=k^{\theta}$, then

$$\frac{\lambda(k+1)-\lambda(k)}{(k+1)-\quad(k)}=\frac{\psi(k+1)(k+1)^{-(1-\theta)}}{(k+1)^{\theta}-k^{\theta}}$$
$$=\frac{\psi(k+1)}{(k+1)((1+\frac{1}{k})^{\theta}-1)})(1+\frac{1}{k})^{\theta}$$
$$=\frac{\psi(k+1)}{(k+1)(\frac{\theta}{k}+o(\frac{1}{k}))}(1+\frac{\theta}{k}+o(\frac{1}{k}))$$
$$=\frac{k\psi(k+1)}{\theta(k+1)(1+o(1))}(1+\frac{\theta}{k}+o(\frac{1}{k}))$$
$$=c(1+o(1)).$$

Obviously, $(k+1)-\quad(k)>0$, for $k\in N_{+}$, $\quad(k)\to\infty$ (as $k\to\infty$), and

$$\lim_{k\to\infty}\frac{\lambda(k+1)-\lambda(k)}{(k+1)-\quad(k)}=c.$$

From the Stolz theorem, $\frac{\lambda(k)}{(k)}\to c$ (as $k\to\infty$) and $\lambda(s)=ck^{\theta}(1+o(1))$.

Therefore,

$$k(\frac{\lambda(k+1)}{\lambda(k)}-1)=k\frac{\lambda(k+1)-\lambda(k)}{\lambda(k)}$$
$$=\frac{\psi(k+1)k(k+1)^{-(1-\theta)}}{ck^{\theta}(1+o(1))}$$
$$=\frac{\psi(k+1)(k+1)^{\theta}}{ck^{\theta}}(1+o(1))$$
$$=\theta(1+o(1))$$

and

$$\frac{\lambda(k+1)}{\lambda(k)}=1+\frac{\theta}{k}(1+o(1)).$$

Let $\xi(k)$ denote

$$\frac{\lambda(k)}{\frac{1-\theta}{\theta}\eta(k)k^{\theta}}$$

and $\overline{\eta}(k)$ denote $\frac{1}{\eta(k)}$,

we have

$$\xi(k)=\frac{\theta}{1-\theta}\lambda(k)\overline{\eta}(k)k^{-\theta} \tag{3.2}$$

and

$$\frac{\xi(k+1)}{\xi(k)}-1=\frac{\lambda(k+1)\overline{\eta}(k+1)}{\lambda(k)\overline{\eta}(k)}(1-\frac{1}{k+1})^{\theta}-1.$$

From Proposition 1.1 we know that $\overline{\eta}(k)$ is the gradually varying function and then

$$\begin{aligned}\frac{\xi(k+1)}{\xi(k)}-1&=(1+\frac{\theta}{k}+o(\frac{1}{k}))(1+o(\frac{1}{k}))(1-\frac{\theta}{k}+o(\frac{1}{k}))-1\\&=\frac{\theta}{k}(1+o(1)+o(\frac{1}{k}))-\frac{\theta}{k}(1+\frac{\theta}{k}+o(\frac{1}{k}))+o(\frac{1}{k})\\&=o(\frac{1}{k}).\end{aligned}$$

From (3.2), $\lambda(k)=\frac{1-\theta}{\theta}\xi(k)\eta(k)k^{\theta}$. Using Corollary 1.1, we conclude that $\varphi(k)=\frac{1-\theta}{\theta}\xi(k)\eta(k)$ varies gradually.

From Corollary 2.1, the fact that $k^{1-\theta}G(k)=\eta(k)\sim\frac{\theta c}{1-\theta}$ and $\eta(k)$ is the gradually varying function implies that $k^{2-\theta}p(k)=\psi(k)\sim\theta c$. This shows that $p(k)=o(\frac{1}{k})$ and $F(k)$ is the gradually varying function (from Theorem 2.2).

Theorem 3.1 $\lambda(k)=F(k)\varphi(k)k^{\theta}$ ($\theta\in(0,1)$), where $F(k)$ and $\varphi(k)$ are gradually varying functions, $\varphi(k)\sim c>0$, and $k\in N_{+}$, if and only if $G(k)=\eta(k)k^{-(1-\theta)}$, where $\eta(k)$ varies gradually, $\eta(k)\sim\frac{\theta c}{1-\theta}$, and if and only if $p(k)=\psi(k)k^{-(2-\theta)}$, where $\psi(k)\sim\theta c$.

Proof. Combining Lemma 3.1, Lemma 3.2, Theorem 2.1 and Corollary 2.1 yields the desired result.

Theorem 3.2 If $F(k)$ varies gradually (at ∞), then

$$N(k)=\overline{\varphi}(k)k^{1-\theta}\ (\theta\in(0,1),\ k\in N_{+}),$$

where $\overline{\varphi}(s)$ varies gradually and $\overline{\varphi}(s) \to \bar{c}$ (as $k \to \infty$), if and only if $p(k) = \psi(k)k^{-(2-\theta)}$, where $\psi(k) \to \theta / \bar{c}$ (as $k \to \infty$).

Proof. $N(k) = \dfrac{k}{E(X \mid X \le k)} = \dfrac{kF(k)}{\lambda(k)} = \overline{\varphi}(k)k^{1-\theta}$, where $F(k)$ and $\overline{\varphi}(k)$ are gradually varying functions. Let $\varphi(k) = \dfrac{1}{\overline{\varphi}(k)}$, then $\lambda(k) = \varphi(k)F(k)k^{\theta}$. From Proposition 2.1, $\varphi(k)$ varies gradually. Using Theorem 3.1 and noting $\varphi(k) \to c = \dfrac{1}{\bar{c}}$ (as $k \to \infty$), we conclude that $p(k) = \psi(k)k^{-(2-\theta)}$, where $\psi(k) \to \theta / \bar{c}$ (as $k \to \infty$). Conversely, from $p(k) = \psi(k)k^{-(2-\theta)}$ with $\psi(k) \to \theta / \bar{c}$, using Theorem 3.1, we get that $\lambda(k) = F(k)\varphi(k)k^{\theta}$, where $\varphi(k)$ varies gradually and $\varphi(k) \sim c = 1/\bar{c}$. Then $N(k) = \dfrac{kF(k)}{\lambda(k)} = \dfrac{1}{\varphi(k)}k^{1-\theta} = \overline{\varphi}(k)k^{1-\theta}$. From Proposition 2.1, $\overline{\varphi}(k)$ varies gradually and $\overline{\varphi}(k) \to \bar{c}$ (as $k \to \infty$).

Conclusions

The main results of the paper may be summarized in the followings.

1. $p(k)$ is the power law distribution on N_+, i.e., $p(k) = \xi(k)k^{-(\alpha+1)}$, where $\lim_{k\to\infty} \xi(k) = c > 0$ and $\alpha > 0$, if and only if $G(k) = \sum_{i \ge k} p(i) = \eta(k)k^{-\alpha}$, where $\lim_{k\to\infty} \eta(k) = \dfrac{c}{\alpha}$ and $\dfrac{\eta(k+1)}{\eta(k)} = 1 + o(\dfrac{1}{k})$.

2. Let k be the number of word occurrences in a text and $N(k)$ be the theoretical number of distinct words in the text. Then, $N(k) = \overline{\varphi}(k)k^{1-\theta}$, where $\theta \in (0,1)$, $\overline{\varphi}(k)$ varies gradually and $\overline{\varphi}(k) \to \bar{c}$ (as $k \to \infty$), if and only if $p(k) = \psi(k)k^{-(2-\theta)}$, where $\psi(k) \to \theta / \bar{c}$ (as $k \to \infty$).

Acknowledgments

This work is supported by NNSFC under Grant No. 61174160. The authors thank three referees for their comments and useful suggestions.

Appendix: An Illustrative Example

This appendix focuses on the importance of the notion of gradual variation in studying power law distributions by showing an illustrative example. It is interesting that we can construct a positive probability $p(k) = \psi(k)k^{-(\alpha+1)}$ on

N_+ ($\alpha > 0$) with the properties (1) there exist reals a and b, $b > a > 0$, such that $\limsup_{k\to\infty} \psi(k) = b$ and $\liminf_{k\to\infty} \psi(k) = a$; and (2) $G(k) = \sum_{i\geq k} p(i) = \eta(k)k^{-\alpha} \sim ck^{-\alpha}$ but $\eta(k)$ does not satisfy the condition of gradual variation.

Let $\eta(k) = 2(1 + \frac{1}{k+1}(-1)^k)$ ($k \in N_+$), then $\eta(k) > 0$ and $\eta(k) \to 2$ (as $k \to \infty$). Let $H(k) = \eta(k)k^{-3}$. From

$$3 + 3\frac{1+(-1)^k}{k} \geq 1 + 2(-1)^{k+1},$$

i.e.,

$$\frac{3}{k}(k + 1 + (-1)^k) \geq 1 + 2(-1)^{k+1},$$

we have

$$\frac{3}{k} \geq \frac{1 + 2(-1)^{k+1}}{k + 1 + (-1)^k} \tag{A.1}$$

Combining $(1 + \frac{1}{k})^3 > 1 + \frac{3}{k}$ and (A.1) yields

$$\frac{1 + 2(-1)^{k+1}}{k + 1 + (-1)^k} + 1 \leq 1 + \frac{3}{k} < (1 + \frac{1}{k})^3. \tag{A.2}$$

From

$$\frac{\eta(k+1)}{\eta(k)} = (1 - \frac{1}{k+2})(\frac{1 + 2(-1)^{k+1}}{k + 1 + (-1)^k} + 1)$$

and (A.2), we have

$$\frac{\eta(k+1)}{\eta(k)} < (1 + \frac{1}{k})^3$$

and $H(k+1) < H(k)$. Let $h(k) = H(k) - H(k+1)$, then

$$\begin{aligned} \sum_{i=1}^{k} h(i) &= \sum_{i=1}^{k} (H(i) - H(i+1)) \\ &= H(1) - H(k+1). \end{aligned}$$

Because $H(1) = 1$, $\lim_{k\to\infty} \sum_{i=1}^{k} h(i) = \sum_{i=1}^{\infty} h(i) = H(1) - \lim_{k\to\infty} H(k+1) = 1$, we conclude that $h(k)$ is the corresponding (right cumulative) distribution function. We replace $h(k)$ and $H(k)$ with $p(k)$ and $G(k)$ respectively.

From $\eta(k) \sim 2$ and

$$\frac{\eta(k+1)}{\eta(k)} = \frac{\eta(k+1)-\eta(k)}{\eta(k)}$$
$$\sim \frac{1}{2}(\eta(k+1)-\eta(k))$$
$$= \frac{1}{2}(-1)^{k+1}\frac{2k+3}{(k+1)(k+2)}$$

it follows that

$$k(\frac{\eta(k+1)}{\eta(k)}-1) \to \begin{cases} 1, & k\ odd, k \to \infty, \\ -1, & k\ even, k \to \infty. \end{cases}$$

Thus, $\eta(k)$ is not the gradually varying function and from Definition 2.3 the corresponding probability $p(k)$ is not the power law distribution.

Consider

$$\frac{\eta(k+1)}{\eta(k)} = 1 + \frac{\eta(k+1)-\eta(k)}{\eta(k)}$$
$$= 1 + (-1)^{k+1}\frac{2k+3}{\eta(k)(k+1)(k+2)}$$

and let $\delta(k) = (-1)^{k+1}\frac{2k+3}{\eta(k)(k+1)(k+2)}$,

we have

$$\frac{G(k+1)}{G(k)} = (1+\delta(k))(1-\frac{1}{k+1})^3$$
$$= (1+\delta(k))(1-\frac{3}{k}+o(\frac{1}{k})) \quad \text{(A.3)}$$
$$= 1-\frac{3}{k}+\delta(k)(1-\frac{3}{k})+o(\frac{1}{k}).$$

Substituting (A.3) into $p(k) = G(k) - G(k+1)$
yields

$$
\begin{aligned}
p(k) &= G(k)\left(\frac{3}{k} - \delta(k)\left(1-\frac{3}{k}\right) + o\left(\frac{1}{k}\right)\right) \\
&= \frac{3}{k} G(k)\left(1 - \frac{k}{3}\delta(k) + o(1)\right) \\
&= 6\left(1 + \frac{1}{k+1}(-1)^k\right)\left(1 - \frac{k}{3}\delta(k) + o(1)\right)k^{-4} \\
&= \psi(k)k^{-4},
\end{aligned}
$$

where $\psi(k) = 6\left(1 + \frac{1}{k+1}(-1)^k\right)\left(1 - \frac{k}{3}\delta(k) + o(1)\right)$. Because

$$
\frac{k}{3}\delta(k) \to \begin{cases} \frac{1}{3}, & k\ odd, k \to \infty, \\ -\frac{1}{3}, & k\ even, k \to \infty, \end{cases}
$$

and

$$
\psi(k) \to \begin{cases} 4, & k\ odd, k \to \infty, \\ 8, & k\ even, k \to \infty, \end{cases}
$$

we conclude that the limit $\lim_{k\to\infty} \psi(k)$ does not exist.

References

Egghe, L. (2007). Untangling Herdan's law and Heaps' law: mathematical and informetric arguments. *Journal of the American Society for Information Science and Technology*, 58, 702-709.

Heaps, H. S. (1978). *Information Retrieval, Computational and Theoretical Aspects*. New York: Academic Press.

Klambauer, G. (1975). *Mathematical analysis*. New York: Macel Dekker.

Lansey, J. C. & Bukiet, B. (2009). Internet search result probabilities: Heaps' law and word associativity. *Journal of Quantitative Linguistics*, 16, 40-66.

Leijenhorst, D. & Weide, T. (2005). A formal derivation of Heaps' law. *Information Science*, 170, 263-272.

Lu, L., Zhang, Z.-K. & Zhou, T. (2010). Zipf's Law Leads to Heaps' Law: Analyzing Their Relation in Finite-Size Systems. *PLoS ONE*, 5, e14139.

Simon, H. (1955). On a class of skew distribution functions. *Biometrika*, 42, 425-440.

Zanette, D. & Montemurro M. (2005). Dynamics of text generation with realistic Zipf's distribution. *Journal of Quantitative Linguistics*, 12, 29-40.

Zhang, Z.-K. et al. (2008). Empirical analysis on a keyword-based semantic system. *European Physical Journal B*, 66, 557-561.

Zipf, G. K. (1936). The Psycho-Biology of Language. An Introduction to Dynamic Philology. London: Routledge.

THE RELATIONSHIP BETWEEN COLLABORATION AND PRODUCTIVITY FOR LONG-TERM INFORMATION SCIENCE RESEARCHERS (RIP)

Jonathan M. Levitt[1] and Mike Thelwall[2]

[1] *J.M.Levitt@wlv.ac.uk and J.Levitt@lboro.ac.uk*
Statistical Cybermetrics Research Group, School of Technology, University of Wolverhampton, Wulfruna Street, WV1 1LY Wolverhampton (UK)

[2] *M.Thelwall@wlv.ac.uk*
Statistical Cybermetrics Research Group, School of Technology, University of Wolverhampton, Wulfruna Street, WV1 1LY Wolverhampton (UK)

Abstract
Over the past decade funding bodies have tended to encourage research collaboration, perhaps because articles by smaller groups of researchers tend to be less highly cited than articles by larger groups. This does not imply that, in general, researchers in smaller groups are less productive since they may produce more articles and hence may receive more citations altogether. This study investigates the relationship between average co-authorship group size and productivity in 2001-2008 for long term researchers, those who authored at least one information science article in both 1998-2001 and 2008-2011. In general the more collaborative researchers were the least productive, supporting previous similar findings for physics. Nevertheless, the most productive information scientists had mean group size of 2-3, suggesting that promoting a small rather than a large degree of research collaboration may be the best strategy for achieving high productivity.

Conference Topic
Collaboration Studies and Network Analysis (Topic 6) and Science Policy and Research Evaluation: Quantitative and Qualitative Approaches (Topic 3).

Introduction
Funding bodies typically promote collaborative research with the implicit or explicit assumption that researchers in collaborative groups are more productive. This assumption may partly stem from collaborative articles being, on average, more highly cited than sole authored articles (e.g., Vogel, 1997; Leta & Chaimovich, 2002; Levitt & Thelwall, 2010). For example, in their investigation of Web of Science articles from 2000 to 2009 with at least one author affiliated to Harvard University, Gazni and Didegah (2011) found a significant positive correlation between the number of co-authors and the number of citations and that the mean citation per article was about one for sole author articles, about two for 2, 3, 4, 5 and 6 co-authors and thereafter rose steadily as the number of co-authors

increased. In an investigation of all articles indexed by the Science Citation Index in 1992, Persson, Glanzel and Danell (2004) found that the mean citation rate, excluding self-citations, increased, on average, by about .58 citations for each additional author. On the same dataset Glanzel, Debackere, Thijs and Schubert found that the self-citation rate for one author articles was 20% and for each collaborative level from 2 to 10 authors was between 25 and 28%, suggesting that self-citations may be partly the cause of the apparent citation advantage of co-authored articles. Glanzel and Danell (2004) also found that for articles published in 1988 and 1998 the mean citation rate of internationally collaborative articles rose more sharply than that for domestic collaborative articles when the number of authors increased. But higher average citation does not imply higher productivity. For example, consider the case of Researcher A who is sole author to one article that received four citations and Researcher B who is sole author to ten articles that in total received thirty citations. Although A has higher average citations than B, A is not more productive than B. Moreover, the productivity of A would be even lower if their article was co-authored and credit for its citations shared with A's co-authors.

A pilot investigation of physics found the relationship between group size and productivity depends on how productivity is measured (Levitt, 2011): when productivity is measured by articles per researcher, productivity decreases as the number of co-authors increases but when productivity is measured by the number of articles per citation band (i.e., clustering articles together with similar citation rates), group sizes of one to two were most productive in the lowest citation band but not in the highest citation bands, suggesting that collaboration may increase the chance of producing the most highly cited research. This pilot investigation investigated research published in a single year and therefore gave an advantage to researchers that publish occasionally since occasional researchers not publishing in the year studies would be excluded altogether. The current paper addresses this limitation by investigating long term researchers and including years in which they did not publish. In addition it investigates the collaboration level variations to see if researchers vary their collaborative style.

Research questions

This study investigates the following research questions:

1. To what extent does the research productivity of long-term researchers vary with the number of collaborative partners (Productivity of groups)?
2. To what extent does the number of collaborative partners of long-term researchers vary (Variation in group size)?

The research questions are addressed for the Information Science & Library Science (IS&LS) Web of Science category with data from articles published in nine years, 2001-2008.

Method

In this study Long-Term researchers are defined as researchers who authored at least one IS&LS article in both 1998-2002 and 2008-11. This study investigates the IS&LS articles by the long-term researchers during 2001-08. The end year of 2008 provides recent data without reducing the citation window to less than 4 years. The set includes not only researchers who published in the start year (2001), but also researchers who published in any of the three years before the start year; and not only researchers who published in the end year (2008), but also researchers who published in any of the three years after the end year. This avoids biasing the sample towards authors that publish every year, who are presumably the most productive. The remaining authors that published at least one article in 2001-2008 are termed 'Occasional' researchers; note that this group contains any long term researcher that did not publish after 2007 or who did not publish an IS&LS article in both 1998-2001 and 2008-2011, due to a career break or due to publishing in a different discipline.

Research productivity was measured by calculating the fractional article contribution and fractional citation contribution in the period assessed. Fractional contributions use the fractional counting system, recommended by Price (19981) and used in several investigations (e.g., Burrell & Rousseau, 1995; Glänzel & De Lange, 2002). In the fractional counting system the credit for a sole author article is the same as the total credit for a collaborative article, but the authors of collaborative articles share equal fractions of the credit for their articles. We decided against measuring contribution using the whole counting system, in which the article and citation credit is irrespective of the number of authors, because the objective is to identify the impact of collaboration on productivity overall in the system.

Bibliographic data on IS&LS articles was obtained from the Social Sciences Citation Index in autumn 2011 and the names of researchers extracted. For each period and each researcher name, the collaboration level of the researcher was obtained by adding the number of authors in the IS&LS articles authored by the researcher in the period and dividing by the number of IS&LS articles published by the researcher in the period. Authors were identified on the basis of their surnames and initials. As the number of Long-Term researchers was less than 1,800, there would have been few cases of different authors to have had the same surname and initials and these should not systematically bias the results.

Findings

Productivity of groups (Question 1)

Tables 1 and 2 compare the productivity of long-term researchers with the productivity of Occasional researchers. In tables 1-4, 'All' refers to all levels of

researchers, '1–2' to the researchers with a collaboration level less than 2, and '8–9' to the researchers with a collaboration level of at least 8 but less than 9.

Table 1: Productivity by collaboration level for Long-Term IS&LS researchers (2001-2008) using the whole number counting system.

Collaboration level	*All*	*1–2*	*2–3*	*3–4*	*4–5*	*5–6*	*6–7*	*7– 8*	*8– 9*
Number of researchers	1,722	20.9%	34.0%	22.6%	10.5%	5.4%	2.2%	2.0%	1.1%
Mean level of collaboration	3.03	1.28	2.31	3.21	4.23	5.20	6.28	7.22	8.16
Articles per researcher	4.81	5.20	5.59	4.62	3.77	3.43	4.55	2.31	2.11
Citations per article	14.65	7.34	16.78	15.09	16.83	15.72	30.04	23.94	23.26

Table 2: Productivity by collaboration level for Occasional IS&LS researchers (2001-2008) using the whole number counting system.

Collaboration level	*All*	*1–2*	*2–3*	*3–4*	*4–5*	*5–6*	*6–7*	*7– 8*	*8– 9*
Number of researchers	21,024	20.3%	26.2%	21.6%	12.4%	6.7%	4.6%	2.6%	1.8%
Mean level of collaboration	3.30	1.08	2.09	3.07	4.05	5.06	6.04	7.04	8.03
Articles per researcher	1.44	1.72	1.49	1.42	1.27	1.30	1.22	1.16	1.17
Citations per article	9.83	3.76	10.41	12.38	11.11	10.87	12.65	15.52	13.53

Again using the whole number counting system, for Long-Term and Occasional researchers the Spearman correlations between the mean collaboration level and articles per researcher were small at 0.016 ($p > 0.05$, n=1,722) and -0.031 ($p < 0.001$, n=21,024) respectively, despite the clear decreasing trends in tables 1 and 2. For the Long-Term and Occasional researchers the Spearman correlations between the mean collaboration level and citations per article were moderate at 0.260 ($p < 0.001$) and 0.265 ($p < 0.001$). Amongst the researchers who published at least one IS&LS article in 2001-2009, fewer than 7.6% were Long-Term. These Long-Term researchers were particularly productive; on average they authored 3.34 times as many articles as the Occasional researchers and their articles on average received 49% more citations each. For both Long-Term and Occasional researchers, a collaboration level of 2–3 is the most common but Occasional researchers have a slightly higher average group size. For both Long-Term and Occasional researchers, for every collaboration level higher than 1–2, the mean citation per article was more than double that for 1–2, although the mean number

of citations per article varied little as the collaboration level increased from 2–3 to 5–6. The picture is different when fractional counting is used, however.

Table 3 indicates that, apart from two years (2004 and 2006), the fractional citation contribution per long-term researcher was lower for the 1–2 collaboration level than for the 2–3 level and thereafter in all years it decreased steadily as the collaboration level increased.

Table 3: Fractional citations per researcher by collaboration levels for Long-Term researchers.

Year	*All*	*1–2*	*2–3*	*3–4*	*4–5*	*5–6*	*6–9*
2001	11.57	14.31	16.16	9.52	6.04	4.85	3.83
2002	15.11	17.79	18.52	14.08	7.30	6.67	1.74
2003	13.33	10.11	18.70	10.87	17.38	6.51	10.27
2004	12.41	15.84	13.63	13.43	6.99	4.98	5.28
2005	9.72	9.98	13.59	9.58	7.43	3.30	4.06
2006	10.38	16.55	13.14	7.13	5.87	7.26	4.33
2007	6.90	9.15	11.49	4.54	3.30	1.99	4.10
2008	4.75	6.42	6.86	3.64	2.95	2.20	1.67

In all apart from two cases (3–4 to 4–5 for 2004; 4–5 to 5–6 for 2002) the fractional article contribution decreased as the collaboration level increased from 1–2 to 4–6 apart from two exceptions (Table 4).

Table 4: Fractional articles per researcher by collaboration level for Long-Term researchers.

Year	*All*	*1–2*	*2–3*	*3–4*	*4–5*	*5–6*	*6–9*
2001	.77	1.48	.77	.51	.36	.29	.22
2002	.95	1.93	.85	.60	.38	.39	.19
2003	.96	1.77	1.03	.59	.53	.27	.27
2004	.86	1.67	.81	.48	.48	.30	.27
2005	.82	1.60	.90	.51	.44	.32	.24
2006	.79	1.51	.92	.59	.40	.27	.27
2007	.82	1.63	1.02	.53	.38	.29	.25
2008	.75	1.47	.82	.53	.38	.31	.25

Table 5 shows the relationship between collaboration level and number of articles authored. In Table 5, the ‘Sample size’ column presents the number of long-term researchers at the collaboration level, and the remaining columns present the percentage of researchers at the collaboration level who authored the number of articles in the column heading. For example, the ‘24’ in the in the ‘1–2’ row and ‘1’ column indicates that 24% of the Long-Term researchers at the collaboration

level of >= 1 and < 2 authored exactly one IS&LS article in 2001-2008. For all collaboration levels in Table 5, over half the researchers authored fewer than 4 articles. The percentage of researchers who authored one article only was substantially lower for the 2–3 collaboration level than for other collaboration levels. There is also a suggestion in the table that lower collaboration levels are associated an increased chance of the highest productivity (larger percentages in the >8 articles column).

Table 5: Percentage of Long-Term researchers by the number of articles published (top row).

Collaboration level	*Sample size*	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*> 8*
1–2	360	24%	22%	12%	11%	6%	4%	4%	6%	11%
2–3	585	17%	16%	18%	10%	8%	6%	4%	3%	16%
3–4	389	27%	17%	15%	12%	5%	5%	4%	2%	13%
4–5	181	26%	19%	12%	13%	11%	5%	4%	2%	8%
5–6	93	30%	19%	15%	9%	5%	7%	7%	3%	5%

Variation in group size (Question 2)

In order to gauge variations in group size, for different collaboration levels the maximum and minimum number of authors are listed in Table 6. The minima and maxima are calculated only on researchers that authored more than one article as, in order for the researcher's group size to vary, the researcher needs to have authored more than one article. For example, the '31%' in the in the '1–2' row and '1' column indicates that 31% of the multi-author Long-Term researchers at the collaboration level of >= 1 and < 2 authored a maximum of one IS&LS article in 2001-2008. Table 5 indicates considerable variation in the level of group size. For example, 9% of the multi-article researchers at the collaboration level of 1–2 authored an article that had at least four authors and 31% of the multi-article researchers at the collaboration level of 5–6 authored an article that had fewer than three authors.

Table 6: Minimum and maximum number of authors expressed as a percentage of all multiple-article authors. Shaded cells report maxima and the un-shaded cells report minima.

Collaboration level	*Sample size*	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*> 8*
1–2	274	31%	44%	16%	4%	2%	0%	0%	1%	2%
2–3	486	47%	14%	45%	22%	11%	3%	3%	1%	2%
3–4	286	21%	47%	13%	35%	27%	10%	6%	2%	6%
4–5	134	9%	34%	40%	5%	25%	31%	17%	5%	16%
5–6	65	12%	19%	25%	34%	5%	22%	22%	19%	39%

Limitations and Discussion

A major limitation is that only one dataset was investigated and the findings might be different for other subjects or databases with different coverage (e.g., Scopus). The findings also depend to some extent on the use of fractional counting. Whilst whole counting would give an unfair advantage to highly collaborative authors, it would have been reasonable to use a counting system that allocated fractional credit to scholars based upon their order in the author list. Moreover, some of the authors, and particularly those of highly collaborative papers, may have been non-scientists, scientists from other fields (e.g., computing researchers constructing software for a project) or even listed for honorary purposes (e.g., department heads) and so the categories may mix core and peripheral researchers in different proportions. Hence direct comparisons between collaboration levels are likely to be to some extent unfair.

In order to obtain some indication of the extent to which the findings depend on the dataset for the same subject, the study was partly repeated with a different dataset, researchers who authored at least one IS&LS article in both 2002-2005 and 2008-2011 (called 'Medium-Term researchers'). This enabled comparison of the fractional citation and article contribution of long-term researchers with that of medium-term researchers. In all cases, apart from two years (2004 and 2006), the fractional citation contribution per Long-Term researcher was lower for the 1–2 collaboration level than for the 2–3 level, and thereafter in all years it decreased steadily as the collaboration level increased; in all years the fractional citation contribution per Medium-Term researcher was lower for the 1–2 collaboration level than for the 2–3 level, but then decreased steadily as the collaboration level increased. In all cases, apart from two cases (3–4 to 4–5 for 2004; 4–5 to 5–6 for 2002), the fractional article contribution per Long-Term researcher decreased as the collaboration level increased from 1–2 to 4–6, whereas in all cases the fractional article contribution per Medium-Term researcher decreased as the collaboration level increased. In conclusion, the similarity in the findings between Medium-Term and Long-Term indicates that findings are unlikely to depend much on the length of period investigated.

Conclusions

As has been found previously for many other data sets, Long Term IS&LS researchers' articles were more highly cited when more collaborative. In agreement with one previous study, however, the productivity of Long Term IS&LS researchers decreases as their level of collaboration increases, both in terms of the number of articles produced and the total number of citations received (using fractional counting in both cases). This confirms the previous finding for physics and, because of the focus on long term researchers, ensures that the results cannot be attributed to the exclusion of less productive researchers that may not author an article every year. Hence, the main contribution of this article is the increased evidence (now for two disciplines) that research

productivity seems to decline overall with the average number of collaborators. Despite this, the optimal average collaboration level seems to be 2-3 rather than 1-2 in IS&LS, suggesting that a small amount of collaboration is the optimal strategy. Finally, there was considerable variation in the sizes of groups in which individual researchers participated so IS&LS researchers are willing to experiment with different styles.

Acknowledgements

This research is supported by the Economic and Social Research Council [grant reference: RES-000-22-4415]. We thank Gertrude Levitt for her very helpful comments on the drafts.

References

Burrell, Q. & Rousseau, R. (1995). Fractional counts for authorship attribution: A numerical study. *Journal of the American Society for Information Science* 46(2), 97–102.

Gazni, A. & Didegah, F. (2011). Investigating different types of research collaboration and citation impact: a case study of Harvard University's publications. *Scientometrics* 87(2), 251-265.

Glänzel W. & De Lange C. (2002). A distributional approach to multinationality measures of international scientific collaboration. *Scientometrics* 54(1), 75-89.

Glanzel, W., Debackere, K., Thijs, B. & Schubert, A. (2006). A concise review on the role of author self-citations in information science, bibliometrics and science policy. *Scientometrics* 67(2) 263-277.

Leta, J. & Chaimovich, H. (2002). Recognition and international collaboration: The Brazilian case. *Scientometrics* 53(3), 325–335.

Levitt J.M. (2011). Preliminary findings on whether it is good value for money to fund larger research groups. *ISSI Newsletter*.

Levitt J.M. & Thelwall M. (2010). Does the higher citation of collaborative research differ from region to region? A case study of Economics. *Scientometrics* 85 (1), 171–183.

Persson, O., Glanzel, W., & Danell, R. (2004). Inflationary bibliometric values: The role of scientific collaboration and the need for relative indicators in evaluative studies. *Scientometrics* 60(3), 421–432.

Price, D.J. de Solla (1981). Multiple Authorship. Science 212(4498), 986.

Vogel, E.E. (1997). Impact factor and international collaboration in Chilean physics: 1987–1994. *Scientometrics* 38(2), 253–263.

RELATIONSHIP BETWEEN DOWNLOADS AND CITATION AND THE INFLUENCE OF LANGUAGE

Vicente P. Guerrero-Bote[1] and Félix Moya-Anegón[2]

[1] *guerrero@unex.es*
Grupo Scimago, Universidad de Extremadura, Departamento de Información y Comunicación, Plazuela Ibn Marwan, 06071 Badajoz (Spain)

[2] *felix.demoya@cchs.csic.es*
Grupo Scimago, CSIC, CCHS, IPP, C/Albasanz, 26-28, 28037 Madrid (Spain)

Abstract
Download indicators represent a great potential due to the high amount of download data that can be collected that can provide a great statistical significance. The relationship between citation and downloads at journal level and the influence of language on it is studied with the data of Scopus (for citation) and ScienceDirect (for downloads).
The results show that the use of downloads as prediction of the citation, is limited, as in the early years is when it obtained less significance. The relationship between downloads and citations is also different in different areas.
In Francophone regions the downloads of English language journals is proportionately greatly reduced with respect to their citation. There seems to be a part of the citation impact of the non-English language journals invisible in Scopus, which make the number of downloads proportionally greater than citations. This has its effect on the lack of correlation between the downloads and citations in the non-English language journals.

Conference Topic
Scientometrics Indicators: Criticism and new developments (Topic 1) and Old and New Data Sources for Scientometric Studies: Coverage, Accuracy and Reliability (Topic 2).

Introduction
The bibliometric indicators used for research evaluation not only take into account quantitative but also qualitative aspects. This is based on the citation of papers included in the main databases (Thomson Reuters Web of Science and Scopus principally) and on the idea that in spite of the different motivations (Brooks, 1985), citations are recognitions of previous works (Moed, 2005a). However, frequently, the application of these bibliometric indicators and these international databases to certain disciplines has been questioned. For some, the bibliometric indicators built from these databases are effective normally in basic science contexts in which research is spread mainly thru scientific journals (Filippo & Fernández, 2002). Different research fields have varying yearly average citation rates (Lundberg, 2007). Bibliometric indicators are almost always lower in areas of Engineering, Social Sciences and Humanities using the ISI data

(Guerrero et al., 2007) and using Scopus data (Lancho-Barrantes, Guerrero-Bote & Moya-Anegón, 2010).

Throughout the scientific literature some authors have pointed out a lack of statistical significance and normalization that could have been originated by a series of causes: the lack of database coverage in certain areas (Braun, Glänzel & Schubert, 2000, Grupo SCImago, 2006, Moya-Anegón et al., 2007), both for the journals and principally for types of documents, and the referencing habits in the different scientific areas (Broadus, 1971; Clemens et al., 1995; Cronin, Snyder & Atkins, 1997; Hargens, 2000; Kyvik, 2003; Lewison, 2001; Lindholm-Romantschuk & Warner, 1996; Nederhof et al., 1989; Nock, 2001; Price, 1970; Small & Crane, 1979; Thompson, 2002).

Since scientific literature is now mostly published and accessed online, a number of initiatives have attempted to measure scientific impact from download log data. The download data allows scientific activity to be observed immediately upon publication, rather than to wait for citations to emerge in the published literature and to be included in citation databases; a process that with average publication delays can easily take several years. Shepherd (2007) and Bollen et al, (2008) propose a Download Impact Factor as journal metrics which consists of average download rates for the articles published in a journal, similar to the citation-based JIF. Bollen et al. (2005, 2008) demonstrate the feasibility of a variety of social network metrics calculated on the basis of download networks extracted from the clickstream information contained in download log data.

Bollen et al, (2009) performed a principal component analysis of the journal rankings produced by 39 measures of scholarly impact that were calculated on the basis of both citation and download log data. Their results indicate that the notion of scientific impact is a multi-dimensional construct that cannot be adequately measured by any single indicator, although some measures are more suitable than others. They observed a greater reliability of download measures possibly caused by the high amount of download data that can be collected.

Although Kurtz et al. (2005) shows how the obsolescence function (Egghe & Rousseau, 2000) of citations and readership follow similar trajectories across time, Schloegl and Gorraiz (2010, 2011) shows that downloads and citations have different obsolescence patterns. Darmoni et al. (2002) and Bollen et al. (2009) show that journal download frequency does not correspond very much to the Impact Factor, although Schloegl and Gorraiz (2011) computed a high correlation at the journal level between citation and download frequencies when using absolute values and a moderate to high correlation when relating usage and citation impact factors. Wan, Hua, Rousseau, and Sun (2010) defined also a download immediacy index.

Although citation indicators are accepted by the international scientific community, they have problems of statistical significance and normalization that could have been originated by the lack of database coverage in certain areas, and the referencing habits in the different scientific areas.

Download indicators represent a great potential due to the high amount of download data that can be collected that can provide a great statistical significance. However, studies indicate that these are only loosely related with indicators based on impact. Would it be possible to use downloads as predictors of the citation?
And, there is no study about the influence of the language in downloads and in its relationship with citation. Are there differences between downloads and citation by language of publication? Is download number by publication languages proportional to the citation one? And does the language have influence in the origin of the citation and the downloads?
Thus our purpose is to study the relationship between citation and downloads at journal level, with the volume of data of ScienceDirect and Scopus, and the influence of language on it. The origin of download will be also studied.

Method and Data

The method that we have applied is to relate the downloads of each journal with the citation of that journal so that correlations between them could easily be found. Not to be very rough, and make a finer comparison, we have compared download counts of downloaded and downloading year with citations to cited year from citing year.
To this goal we have used the download data from ScienceDirect and the citation data from Scopus.
To set the language differences, we have studied these parameters for non-English journals in ScienceDirect. More particularly those having more than 95% of the papers in French (15), German (4) or Spanish (4) in the period 2003-2011.
We have also defined a control group of English journals in ScienceDirect, so as to establish the differences between the non-English and English journals. For every non-English journal, at least one English journal present in both databases, belonging to the same Specific Subject Area and with similar number of papers published was selected as control journal, up to 33 control journals.
To go deeper into this, we have compared the geographic origin of both the download with the citation of both groups.
Not all journals publish papers every year. There are only 8 non-English journals (French) with papers every year and 14 control journals. The rest of journals begin or are incorporated during the period, because of that they have no papers in the first years. However, there are three exceptions, one French journal with no paper the last year of the period and two control journals with no papers in the last two years.
The majority of the journals are concentrated in the Subject Area of Medicine, where all the German and Spanish journals are added and the majority of the French. Two other French journals are from "Pharmacology, Toxicology and Pharmaceutics", and three from Psychology (although one of the latter also is assigned to Medicine). The majority of the control journals are included also in “Medicine” (27) (2 in "Pharmacology, Toxicology and Pharmaceutics" and 6 in

Psychology), however, many other subject areas appear because of the addition of journals to multiple Subject Areas.

In the data from ScienceDirect supplied by Elsevier, each paper, regardless of documental type, has two dates, the online date and the publishing date. It is usual at present to publish a paper online before in the journal issue, a way to take advantage of the "Early View" effect (Moed, 2005b; Craig et al., 2007; Davis et al., 2008). We have calculated the difference between these two dates (publication date minus online date expressed in days), and it can be observed (in table 1) that in the first part of the period such difference is negative (they were published online some time after they were published in the issue). It can be observed also how during the period the difference closes to cero and becomes positive at the end of the period. That may be caused by a retrospective incorporation to ScienceDirect. For example, Masson journals started in ScienceDirect in 2010/2011, while they were already quite far in their volume numbering. E.g., Masson published volumes 1-10, and first Elsevier processed issue is volume 11 in 2010. What ScienceDirect then does is back-capture volumes 1-10 and add those to ScienceDirect. The on-line dates are then the dates the back-captured articles are added to ScienceDirect. That means that the publication/cover dates are older than the on-line dates, which gives those weird minus figures in the tables. This especially happened with Spanish titles, and also with French ones, although many French journals were already in ScienceDirect for a long time so that the effect is less. German titles from Urban&Fischer show the same patterns, although less since Elsevier acquired U&F much earlier than the Spanish publications.

The types of documents of the data provided by Elsevier ScienceDirect that accumulate more than 5% of downloads which have more than 500 downloads per paper and which accumulate a percentage of downloads superior to its percentage of papers are Review Article, Short Survey, Full length article and Short Comunication. The other types of documents do not involve major scientific contributions. Therefore in this paper we focus on these four document types from ScienceDirect as primary production.

In Scopus, documental typology is slightly different. The three types that accumulate more than 2% of the citation and more than 5 citations per paper on average are Reviews, Articles and Conference Papers. However, while this is true in this Journal Set, in general in Scopus, the Short Surveys accumulate a citation similar to the Conference Papers, so that we included it in this study, along with the above three as primary documents.

The records from ScienceDirect are 79,363, while the records from Scopus are 43,914. The divergence is mainly because Scopus covers all items except types: conference/meeting abstracts and book reviews. Specifically, the abstracts represent over 38% of the records of ScienceDirect.

Table 1: Average difference between the online date and the publication date (publication date minus online date) expressed in days. They are grouped by the year of publication in the issue.

Journal	*Language*	*Ndocc*	*2003*	*2004*	*2005*	*2006*	*2007*	*2008*	*2009*	*2010*	*2011*
Acute Pain	English	132	15.4	-68.1	21.1	37.1	43.6	37.2	54.7		
Addictive Behaviors	English	1819	190.7	106.9	177.9	235.7	224.1	140.5	129.1	127.4	122.8
Alzheimer's & Dementia	English	290			-76.0	-5.5	2.2	13.2	0.6	23.8	13.8
Asian Journal of Psychiatry	English	137						3.8	6.2	13.8	14.4
Biomedical and Environmental Sciences	English	322						-79.5	-75.4	-64.3	-66.2
Children and Youth Services Review	English	1150	6.9	71.1	145.1	233.5	111.5	178.8	153.8	139.3	138.8
Clinical Microbiology Newsletter	English	341	-95.6	-16.5	-26.5	-1.5	-1.1	-1.8	5.5	-2.2	5.5
Contraception	English	1337	-11.5	9.4	39.4	85.5	50.2	51.0	111.6	117.2	172.1
Diagnostic Microbiology and Infectious Disease	English	1789	50.9	10.4	26.7	78.5	77.9	81.4	43.0	63.1	38.4
Early Human Development	English	1014	35.3	44.8	51.7	94.4	170.9	137.8	68.3	18.4	53.0
e-SPEN, the European e-Journal of Clinical Nutrition and Metabolism	English	178					4.6	66.1	47.9	39.2	36.9
European Journal of Integrative Medicine	English	67							11.6	35.4	6.3
European Journal of Pharmaceutical Sciences	English	1368	10.5	41.4	55.4	97.0	90.0	79.3	72.5	68.3	51.2
EXPLORE: The Journal of Science and Healing	English	460			-12.3	-15.1	-24.9	-3.5	-5.4	-7.8	-4.2
Forensic Science International Supplement Series	English	30							53.6		
General Hospital Psychiatry	English	766	-17.4	-18.3	-10.5	-2.4	-4.1	22.4	98.6	102.0	48.4
International Journal of Drug Policy	English	450	1.1	49.7	23.2	35.8	142.5	214.7	256.9	195.4	62.0
International Journal of Pediatric Otorhinolaryngology Extra	English	366				51.5	75.7	123.0	228.9	359.4	309.8
Japanese Dental Science Review	English	49						33.8	54.3	152.4	139.1
Journal of Adolescent Health	English	1598	8.7	13.0	9.4	35.2	53.3	84.5	114.8	121.4	139.5
Journal of Cardiology Cases	English	175								127.9	70.8
Journal of Hepatology	English	2325	35.3	69.4	90.9	99.3	97.3	76.2	81.0	84.6	193.4
Journal of Medical Colleges of PLA	English	259					-90.6	-39.3	-70.7	-64.5	-64.5
Journal of Pediatric Urology	English	670			117.9	197.2	200.5	155.6	143.3	212.3	199.6
Journal of the American Academy of Child & Adolescent Psychiatry	English	1278	-2327.0	-1970.0	-1595.0	-1238.6	-864.5	-466.4	-94.0	33.7	32.8
Mental Health and Physical Activity	English	50						38.2	109.5	103.2	122.4
Microbial Pathogenesis	English	711	22.9	42.2	13.2	35.7	84.0	118.2	70.4	100.3	95.5
Nanomedicine: Nanotechnology, Biology and Medicine	English	403			-53.1	-13.8	5.5	76.4	182.5	198.9	194.4
Progress in Lipid Research	English	208	80.9	87.8	3.7	80.3	80.4	109.2	76.2	144.4	105.0
Research in Social and Administrative Pharmacy	English	228			-26.1	-8.5	-19.4	25.5	143.2	213.7	331.3
Sexologies	English	222				-32.3	16.5	76.2	38.4	88.9	83.5
Surgical Pathology Clinics	English	131						-5.0	-12.5	-58.6	-4.5
Taiwanese Journal of Obstetrics and Gynecology	English	598		-1812.7	-1460.6	-1089.9	-415.2	-109.1	-48.3	-54.2	-43.1
Actualités Pharmaceutiques	French	461						-238.7	-74.2	-67.0	-112.0
Actualités Pharmaceutiques Hospitalières	French	183			-1290.0	-952.6	-577.7	-260.0	-85.6	-153.6	-135.4
Annales Médico-psychologiques, revue psychiatrique	French	957	-29.3	20.6	37.1	85.5	104.0	224.9	85.9	55.2	81.0
Archives de Pédiatrie	French	2732	-102.0	26.4	45.5	33.6	17.8	-7.3	24.7	-4.5	10.3
Gynécologie Obstétrique & Fertilité	French	1206	-53.3	19.3	4.5	-0.8	-41.8	-3.0	-2.2	2.7	14.0
Journal de Pédiatrie et de Puériculture	French	367	-130.2	54.9	30.4	31.5	15.9	22.4	29.6	48.8	59.0
La Revue de Médecine Interne	French	1573	-248.9	-20.8	45.6	81.5	87.0	139.7	144.8	83.2	151.0
La Revue de Médecine Légale	French	29								-16.8	8.6
L'Encéphale	French	968		-1239.1	-862.3	-428.0	-127.9	46.5	88.2	108.5	87.4
Médecine & Droit	French	224	-147.5	-80.8	-42.2	-3.5	22.1	4.0	16.9	14.9	37.8
Neuropsychiatrie de l'Enfance et de l'Adolescence	French	613	-36.2	-24.3	-43.0	-6.5	16.8	66.4	112.4	173.7	201.1
Nutrition Clinique et Métabolisme	French	273	-18.6	9.1	17.0	-62.1	-58.1	-15.9	2.0	9.7	7.8
Pratiques Psychologiques	French	244		9.3	14.9	36.2	53.8	62.1	245.7	329.9	472.2
Psychologie Française	French	202		14.6	81.6	80.0	145.8	115.9	111.0	49.0	33.8
Réanimation	French	132	-25.3	-2.9	-14.1	-6.7	-74.4	-39.5	65.5	-35.7	
Das Neurophysiologie-Labor	German	63					-2.6	5.6	153.5	58.6	59.5
Krankenhaus-Hygiene + Infektionsverhütung	German	105					-5.1	19.9	1.0	30.7	21.0
Osteopathische Medizin	German	70						-151.1	-58.2	-62.3	-30.4
Public Health Forum	German	282					23.6	10.4	44.8	38.1	38.2
Cardiocore	Spanish	90								-13.0	83.6
Revista de Psiquiatría y Salud Mental	Spanish	69						-20.0	-116.0	-62.0	-41.2
Revista Española de Patología	Spanish	243				-1448.7	-1081.4	-706.7	-348.1	-20.8	61.3
Revista Internacional de Acupuntura	Spanish	171					-416.8	-197.1	-68.9	-113.4	-147.5

It may seem a bit inconsistent that one documental type from Scopus considered primary production are "Conference Papers", while the type "Conference" from ScienceDirect has not been considered. However, the percentage involved is quite small, and the percentage of downloads which accumulates is even lower, which means that the number of downloads per papers is below average. And at the same time the "Conference Papers" from Scopus are included primarily as "Full

Length Articles" and only less than 5% of them are listed as "Conference" from ScienceDirect, because ScienceDirect assign 'Full Length Article' to full scientific papers in Conference issues. Then, though "Conference Papers" of Scopus represents the greater part of "Conference" of ScienceDirect, they do not get a large number of downloads.

Results and discussion

In Figure 1, the average of citation from primary documents considered have been represented per Scopus documental types and age. Unlike other similar representations in this case they have not been made per calendar year, but the time difference between the citing and cited document has been considered. That is, for instance to calculate the citation average in the eighth year, only the papers with a minimum of 8 years have been considered, and the average was calculated with citation aged between 7 and 8 years. As in Scopus only pre-2012 citation can be considered almost complete, to calculate the citation average of 8 years only papers published in 2003 were considered since they are the only ones having at least eight full years to receive citations. To calculate the citation average of seven years papers published in 2003 and 2004 were considered since they are the only ones having at least 7 full years to receive citations. And so on. This means that the data for lesser age are statistically more significant because they have been computed with larger datasets.

The citation maximum for Reviews is obtained at 3 years. For articles, the citation for the third and fourth years is very similar, and the fall is much slower, the citation in the seventh year exceeds even that of the second year.

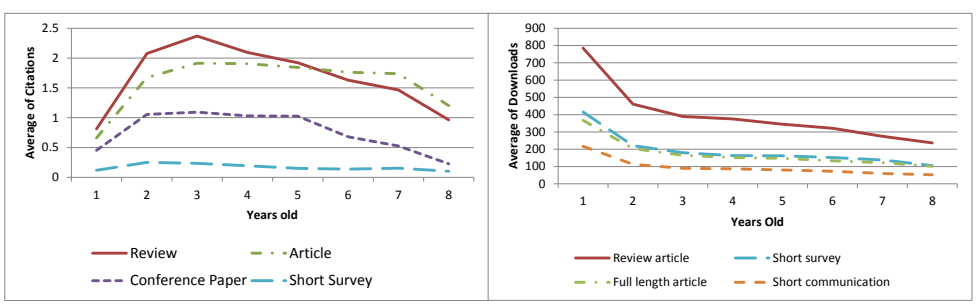


Figure 1: Average of primary Citations per Scopus document type by age in years of the Journal Set and Averages of Downloads of the main Science Direct document types per Science Direct paper by years of difference with respect to the online publication date.

In the second part of Figure 1, a similar representation has been made but with regard to the downloads. In this case, the online date has been used as reference. Similarly, the number of years shown on the horizontal axis refers to the difference between the date of download and online publication date. To calculate each average, only those papers that have the corresponding full annuity to be

downloaded have been considered. In 8, only those downloads in which the difference between the download and the online date is between 7 and 8 years have been considered.
In this case all the curves are monotonic decreasing, while in the previous figure as a result of the time required for citation from the date in which the cited paper is published until the citing paper is made and published.
In the case of downloads, the diffusion made when the paper is published online and the large number of downloads that results from novelty are very evident. Also there is a greater difference between Reviews and Articles.
Figure 2 shows the citation from primary papers toward primary papers by Subject Areas. The way of computation is similar to the previous one. The peak in the seventh year of Pharmacology, Toxicology and Pharmaceutics is striking, however it is because in 2005 two of the four journals of the subject area enter, which lower significantly the citation average, except for that incidence the curves are quite similar.
The second part of Figure 2 similarly shows download averages by Subject areas and years. As in the first part, irregularities in the curve of Pharmacology, Toxicology and Pharmaceutics are observed, following the incorporation of the four journals assigned to it at different times.
Also striking is the difference in the order of the Subject Areas, while Medicine is the one with the highest average of citations, it is the one that has the lowest downloads. Psychology while always behind in citations, is always ahead in downloads. This may, once again, be indicative of different patterns in different areas.

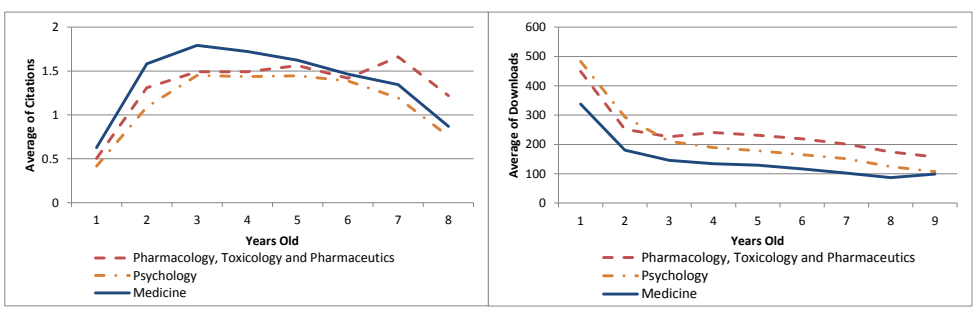


Figure 2: Average of primary citation toward primary papers, and average of downloads of primary papers by Subject Area (only the three original Subject Areas).

Table 2 shows the correlation between averages of downloads and averages of citations by journals, year of publication and of citation or download "age" using the same calculation method as above. The only difference is that, in order to allow both data to be comparable, for download "age", the date of publication of the paper in the journal has been used instead of online publication date as above. In columns the age of the citation and in rows the age of the downloads. The last

column and row correspond to the sum of all averages of citation/downloads as an average of citation/downloads of up to 8 years old. There is a certain time delay between download and citation: if an author downloads an article, he must first read it, include it a new paper he is writing, and that paper must be published. All this may take 1-2 years, sometimes even more, depending upon journal and perhaps field.

Table 2: Correlations between averages of downloads and averages of citations by journals, year of publication and of citation or download "age". In columns the age of the citation and in rows the age of the downloads. The last column/row correspond to the sum of all averages of citation/downloads.

DY\CY	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*Σ*
1	0.77	0.78	0.82	0.85	0.86	0.88	0.93	0.94	0.51
2	0.71	0.75	0.79	0.84	0.87	0.89	0.93	0.94	0.6
3	0.66	0.71	0.76	0.79	0.83	0.85	0.92	0.93	0.63
4	0.63	0.69	0.74	0.77	0.78	0.81	0.86	0.92	0.67
5	0.64	0.68	0.73	0.75	0.76	0.75	0.85	0.89	0.7
6	0.61	0.66	0.7	0.72	0.75	0.76	0.81	0.9	0.71
7	0.73	0.73	0.76	0.77	0.8	0.81	0.82	0.77	0.78
8	0.72	0.72	0.74	0.78	0.8	0.8	0.84	0.82	0.79
Σ	0.65	0.71	0.76	0.77	0.79	0.79	0.83	0.82	0.73

Table 2 shows that the highest correlation is between the number of downloads of one year of difference and the citation 8 years of difference. The average correlation between the number of downloads and the citation of the same age is 0.78 (the diagonal), between the number of downloads and the citation of two more years of difference is 0.84 and between the citation and the number of downloads of 2 years more of difference 0.73. These results are consistent with the idea that there is a certain time delay between download and citation.

Table 3 shows the correlations between downloads and citations of two more years of age separated by groups with levels of statistical significance. The correlations are significant and positive for the total and for the control set. However, they are lower (in some cases even they may be slightly negative) and of little significance in the case of non-English language journals.

Table 4 has been made with the same data as table 3, but in this case columns of downloads have been correlated with the column that sums the averages of citation up to 8 years of age. With this you can see which are the most significant downloads when predicting total citations obtained by each journal. In this case we can see that none of the correlations of the non-English language journals are statistically significant at a level of $\alpha = 0.05$. The most significant is the third year with a correlation below 0.2.

Table 3: Correlations between averages of downloads and averages of citations by journals, year of publication and of citation or download "age". The citation "age" is two years more than download "age". The correlations have been separated by language of publication.

		1->3	*2->4*	*3->5*	*4->6*	*5->7*	*6->8*	*Σ*
Total	*r*	0.82	0.84	0.83	0.81	0.85	0.90	0.73
	α	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01
English	*r*	0.80	0.86	0.83	0.82	0.87	0.90	0.71
	α	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01
Non-English	*r*	0.29	0.33	0.21	-0.03	-0.31	0.36	0.43
	α	0.01	0.02	0.16	0.89	0.18	0.34	<0.01

Table 4: Correlations between averages of downloads and the sum of the averages of citations by journals, year of publication and of citation or download "age". The correlations have been separated by language of publication.

		1	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*Σ*
Total	*r*	0.51	0.60	0.63	0.67	0.70	0.71	0.78	0.79	0.73
	α	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01
English	*r*	0.45	0.56	0.60	0.64	0.68	0.69	0.77	0.74	0.71
	α	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01
Non-English	*r*	0.07	0.15	0.20	0.12	0.04	-0.005	-0.09	0.27	0.43
	α	0.46	0.13	0.08	0.34	0.78	0.98	0.71	0.48	<0.01

The correlations of the control journals are statistically significant and positive. The highest correlation is obtained in the seventh year.

However, in many cases, only the citation obtained in the first three years is taken into account, which is why we have created Table 5 which shows the correlation of the averages of downloads of different ages with the sum of the averages of citation of the first three years. In this case the correlation increases slightly in the early years in cases of non-English journals although the correlation is still quite low. In the case of the control journals, all the correlations have a high level of statistical significance, and the maximum value is obtained in the downloads of seven years of "age", although the first three years rise steeply with respect to the correlations in Table 4.

For the study and comparison of the origin of the downloads and the citation, two tables were generated by countries with the number of citations and downloads of each country to the control journals of English language in one column, to the French-language journals in another, to the German language journals in another and to the Spanish language journal in the fourth. We have also calculated other columns with the total number of citations and downloads and with citations and downloads of the three groups of journals to study (the French language journals,

the German language journals and Spanish language journals). The countries in this table have been ordered by the scientific production in the period. From them we have kept the data of the 50 most productive countries, which are those with more than 25,000 papers in the period 2003-2011.

Table 5: Correlations between averages of downloads and the sum of the averages of citations of up to 3 years old by journal, year of publication and of citation or download "age". The correlations have been separated by language of publication.

		1	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*Σ*
Total	*r*	0.66	0.74	0.73	0.72	0.70	0.68	0.75	0.74	0.75
	α	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01
English	*r*	0.61	0.72	0.71	0.70	0.68	0.65	0.73	0.67	0.73
	α	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01
Non-English	*r*	0.24	0.33	0.33	0.20	0.12	0.09	-0.01	0.27	0.54
	α	<0.01	<0.01	<0.01	0.12	0.44	0.63	0.98	0.49	<0.01

By correlating the columns, correlations higher than 0.98, statistically significant at $\alpha = 0.01$ level were found between downloads and citations to the same type of journals. The downloads and citations of the control journals have correlations higher than 0.93, statistically significant at $\alpha = 0.01$ level with the scientific production and the total sum of downloads and citations (to the control journals and non-English language journals) while downloads and citations of non-English language journals do not correlate significantly, either by language or as a set (only two correlations are significant at $\alpha = 0.01$, which are slightly less than 0.5, both with the total sum of downloads in each country, one of them is of the citations to French journals and the other of the citations to non-English language journals).

As Table 6 shows, the countries with the highest percentage of downloads of control journals (relative to the total number of downloads thereof) are USA, China and UK. The following country is Canada, although it has a higher percentage of downloads of the French journals. The countries with the highest percentage of the French Journals downloads are France, Tunisia, Canada, Algeria and Belgium all Francophones, although Tunisia and Algeria are not among the 50 most productive. Germany, Switzerland and Austria are for German Journals. Switzerland also has a high percentage of downloads from the French Journals. Spain, Brazil, Argentina and Uruguay are for Spanish Journals, the last two at a great distance, and Brazil not being a Spanish speaking country. If we compare these percentages of downloads with the percentage of total downloads, we see that the three countries with the highest percentage of Control Journals have a slightly higher percentage of them than the total, the ratio is slightly greater than unity. In the case of the French Journals, the ratio becomes slightly larger. Curiously, Tunisia and Algeria have the highest ratio. This ratio continues

to increase in the case of the German Journals and especially the Spanish Journals indicating that these downloads are more concentrated in those countries.

Table 6: The Highest Percentages of Downloads of the Control (%CD), French (%FD), German (%GD) and Spanish (%SD) journal group with respect to the total number of downloads of each group, ratio of these percentages with respect to the Download percentage of each country (%TD) and similar citation ratios.

Country	*%CD*	*%FD*	*%GD*	*%SD*	*%CD/%TD*	*%FD/%TD*	*%GD/%TD*	*%SD/%TD*	*%CC/%TC*	*%FC/%TC*	*%GC/%TC*	*%SC/%TC*
United States	34.41%	4.86%	4.74%	0.46%	1.242	0.175	0.171	0.017	1.060	0.229	0.019	0.119
United Kingdom	8.66%	1.72%	2.24%	0.17%	1.222	0.243	0.316	0.024	1.041	0.474	0.186	0.197
China	6.13%	1.83%	4.68%	0.34%	1.188	0.354	0.906	0.066	1.058	0.256	0	0
France	2.70%	48.08%	1.06%	0.01%	0.211	3.752	0.083	0.001	0.599	6.305	0.091	0
Tunisia	0.27%	10.63%	0.10%	0.01%	0.106	4.115	0.038	0.004	0.331	9.842	0	0
Canada	4.53%	5.49%	0.47%	0.14%	0.958	1.161	0.098	0.029	1.029	0.634	0.430	0
Algeria	0.10%	4.31%	0.05%	0.00%	0.093	4.161	0.052	0	0.539	7.107	0	0
Belgium	0.90%	3.25%	0.86%	0.00%	0.633	2.282	0.608	0	0.916	2.118	0	0.299
Germany	1.91%	0.50%	55.72%	0.04%	1.107	0.288	32.320	0.025	1.029	0.532	14.633	0
Switzerland	1.66%	3.04%	6.50%	0.03%	0.840	1.535	3.287	0.017	0.972	1.380	0.674	0
Austria	0.24%	0.03%	5.40%	0.00%	1.169	0.142	26.159	0	1.028	0.621	3.088	0
Spain	1.80%	1.77%	0.51%	78.59%	0.956	0.939	0.269	41.707	0.979	1.110	0	17.820
Brazil	2.02%	1.15%	0.40%	16.23%	1.098	0.626	0.217	8.803	1.008	0.903	0	1.428
Argentina	0.20%	0.04%	0.01%	1.44%	1.208	0.246	0.035	8.712	1.012	0.844	0	1.196
Uruguay	0.01%	0.02%	0.00%	0.92%	0.749	1.555	0.152	60.521	0.938	1.839	0	0

If you calculate a similar ratio to citation, we can see that the control journals have lower ratios (that of downloads) in the case of USA, UK and China and higher in the remaining cases. The French Journals have higher citation ratios in most of the countries shown. The German Journals only in France and Canada. The Spanish Journals only in the case of USA, UK and Belgium.

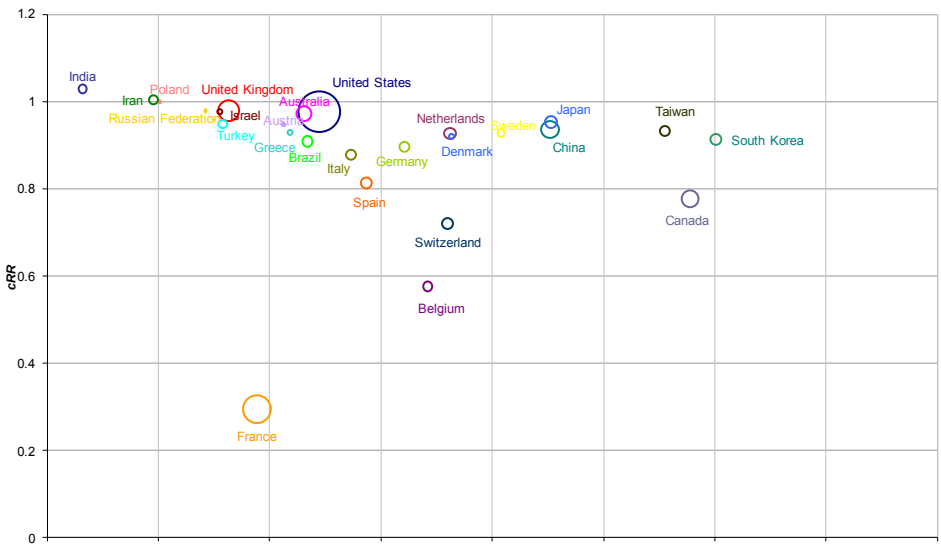


Figure 3: Ratio of downloads with respect to citations of the journals of control (cRR) against the French, German and Spanish language journals (eRR). The size is proportional to the number of total national downloads. The 27 countries with the highest scientific production are represented.

Some others relative columns have also been calculated per country, as the ratio of downloads with respect to citations:

$$gRR_{country} = \frac{\frac{gd_{country}}{td_{country}}}{\frac{gc_{country}}{tc_{country}}}$$

Where *d* are downloads, of a journal group (*gd*) or in total (*td*), and *c* are citations to a journal group (*gc*) or in total (*tc*). The measurements calculated in this way for each country were: *cRR* (ratio of downloads of the control journals with respect to its citations ratio), *fRR* (ratio of downloads of the French journals with respect to its citations ratio), *gRR* (ratio of downloads of the German journals with respect to its citations ratio), *sRR* (ratio of downloads of the Spanish journals with respect to its citations ratio), *eRR* (ratio of downloads of the French, German and Spanish journals with respect to its citations ratio).

Regarding download ratios with respect to the citations, we found that the control journals (*cRR*) average in the top 50 countries of 0.93 with a standard deviation of 0.12. While the group consisting of French, German and Spanish language journals (*eRR*) has an average of 2.35 with a standard deviation of 1.36 (mean difference significant at the $\alpha = 0.01$). This means that in these countries the control journals are cited in greater proportion to the downloading, while the French, German and Spanish languages journals are downloaded twice with respect to its citation. This can be seen in figure 3.
As you can see countries with the lowest downloaded papers of the control journals with respect to the citing, have a francophone link. This effect does not occur in the case of Spanish or German, but we must take into account that the German and Spanish journals studied are very few, and some of them seem to have been included on ScienceDirect retrospectively.

Conclusions

The number of papers from Scopus and ScienceDirect is different, because the first includes all items except types: conference/meeting abstracts and book reviews. The divergence is mainly because of conference/meeting abstracts.
The set of journals in German and Spanish language is not very significant in order to find separate conclusions.
The citation and download curves with respect to time are different. The time required for a paper to be cited can be seen in the citation curves and the effect of novelty in download curves. The proportional difference between the downloads received by the reviews and other types of documents increases with respect to the citation.bi
The order of the Subject Areas in average citation does not match the order in average download. This leads to different patterns in different areas i.e. researchers in different areas cite proportionally differently with respect to what they read.
There are statistically significant correlations between the downloads and citations for journals and years, but these are greatly reduced in both value and statistical significance in the case of non-English language journals. Some influence on these results can have the late incorporation of these journals to ScienceDirect.
In the control journals, at first there is a novelty effect that makes many downloads occur that do not result in citations. This may be the reason why the first year is the one which obtain lower correlations. Interestingly the highest correlations are those of the sixth or seventh year of age, which may correspond to when researchers are looking for a particular paper probably redirected by a citation.
All this makes the use of downloads as prediction of the citation, limited, as in the early years is when it obtained less significance. In no case thus does it reach the correlation between the citation of the first three years with the citation total

(0.91). This circumstance is even greater in the case of non-English language journals.
The 50 most productive countries download the control journals proportional slightly less than they cited them. However the non-English journals to study are downloaded proportionately more than twice what they are cited. This may be due to the fact that a part of the citation impact of the non-English journals is invisible in Scopus because those who download the papers, also cite them in articles published in journals that are not processed for Scopus.
In these 50 most productive countries, there is an association between the proportional citation or downloads of control journals with the ratio between downloads and citation of them. This means that those which frequently proportionally download or cited the control journals, download them proportionately more with respect what they cite them. This same effect does not occur in the non-English journals to study.
In Francophone regions it is observed how the download of control journals is proportionately greatly reduced with respect to their citation. In the case of German and Spanish language, the study is not very significant because the number of journals is very small, some of which have been loaded into ScienceDirect retrospectively.
Definitely there seems to be a part of the citation impact of the non-English language journals invisible in Scopus, which make the number of downloads proportionally greater than citations. This has its effect on the lack of correlation between the downloads and citations in the non-English journals, which means that the downloads can hardly be used to predict the citation.

Acknowledgments

This work was granted by Elsevier as part of the Elsevier Bibliometric Research Program (EBRP) and financed by the Junta de Extremadura, Consejería de Empleo, Empresa e Innovación and the Fondo Social Europeo as part of the research group grant GR10019.

References

Bollen, J., Van de Sompel, H. and Rodriguez, M.A. (2008). Towards usage-based impact metrics: First results from the MESUR project. In *Joint Conference on Digital Libraries (JCDL2006)*, Pittsburgh, PA, June 2008.

Bollen, J., Van de Sompel, H., Hagberg, A, and Chute, R. (2009). A principal component analysis of 39 scientific impact measures. *PLoS ONE*, 4(6): e6022. doi:10.1371/journal.pone.0006022.

Bollen, J., Van de Sompel, H., Smith, J. and Luce, R. (2005). Toward alternative metrics of journal impact: a comparison of download and citation data. *Information Processing and Management*, 41(6):1419-1440.

Braun, T., Glänzel, W., Schubert, A. (2000). How balanced is the Science Citation Index's journal coverage? A preliminary overview of macrolevel statistical data. In: Cronin, B; Barsky Atkins, H (eds.). *The Web of knowledge,*

a festschrift in honor of Eugene Garfield. Canada: American Society of Information Science, 2000, pp. 251–277.
Broadus, R. N. (1971). The literature of the social sciences: a survey of citation studies. *International Social Sciences Journal*, 23: 236–243.
Brooks, T.A. (1985). Private acts and public objects: an investigation of citer motivations. *Journal of the American Society for Information Science*, 36(4): 223-229.
Clemens, E. S., Powell, W. W., Mcilwaine, K., Okamoto, D. (1995). Careers in print: Books, journals, and scholarly reputations. *American Journal of Sociology*, 101: 433–494.
Craig, I., Plume, A., McVeigh, M., Pringle, J., Amin, M. (2007). Do open access articles have greater citation impact? A critical review of the literature. *Journal of Informetrics*, 1, 239-48.
Cronin, B., Snyder, H., Atkins, H. (1997). Comparative citation rankings of authors in monographic and journal literature: a study of sociology. *Journal of Documentation*, 53: 263–273.
Darmoni, S. J., Roussel, F., Benichou, J., Faure, G. C., Thirion, B., & Pinhas, N. (2000). Reading factor as a credible alternative to impact factor: a preliminary study. *Technol. Health Care*, 8 (3-4), 174–175.
Davis, Philip M., Bruce V. Lewenstein, Daniel H. Simon, James G. Booth, and Matthew Connolly. (2008). Open Access Publishing, Article Downloads, and Citations: Randomized Controlled Trial. *British Medical Journal*, 337: 331-345.
Egghe, L., & Rousseau, R. (2000). Aging, obsolescence, impact, growth, and utilization: Definitions and relations. *Journal of the American Society for Information Science*, 51 (11), 1004–1017.
Filippo, D., Fernández, M.T. (2002). Bibliometría: importancia de los indicadores bibliométricos. In: *El estado de la ciencia*. p. 69-76. Red Iberoamericana de Indicadores de Ciencia y Tecnología (RICYT).
Grupo SCImago (2006). Análisis de la cobertura de la base de datos Scopus. *El profesional de la información*, Vol.15, nº2: 144-145.
Guerrero-Bote, V. P., Zapico-Alonso, F., Espinosa-Calvo, M. E., Gómez-Crisóstomo, R., & Moya-Anegón, F. (2007). The Iceberg Hypothesis: Import-Export of Knowledge between scientific subject categories. *Scientometrics*, 71(3): 423-441.
Hargens, L. L. (2000). Using the literature: reference networks, reference contexts, and the social structure of scholarship. *American Sociological Review*, 65 : 846–865.
Kurtz, M. J., Eichhorn, G., Accomazzi, A., Grant, C. S., Demleitner, M., & Murray, S. S. (2005). The bibliometric properties of article readership information. *Journal of the American Society for Information Science and Technology*, 56: 111-28.
Kyvik, S.(2003). Changing trends in publishing behaviour among university faculty, 1980–2000. *Scientometrics*, 58: 35–48.

Lancho-Barrantes, B.S., Guerrero-Bote, V.P. , Moya-Anegón, F. (2010). The Iceberg Hypothesis revisited. *Scientometrics*, 85: 443-461.
Lewison, G. (2001), Evaluation of books as research outputs in history of medicine. *Research Evaluation*, 10 : 89–95.
Lindholm-Romantschuk, Y., Warner, J. (1996). The role of monographs in scholarly communication: an empirical study of philosophy, sociology and economics. *Journal of Documentation*, 54 : 389–404.
Lundberg, J. (2007). Lifting the crown—citation z-score. *Journal of Informetrics*, 1, 145–154 .
Moed, H.F. (2005a). *Citation Analysis in research evaluation*. Dordrecht; Springer, p. 346.
Moed, H.F. (2005b). Statistical relationships between downloads and citations at the level of individual documents within a single journal. *Journal of the American Society for Information Science and Technology*, 56, 1088-97.
Moya-Anegón, F., Chinchilla-Rodríguez, Z., Vargas-Quesada, B., Corera-Álvarez, E., Muñoz-Fernández, F. J., González-Molina, A., et al. (2007). Coverage analysis of Scopus: a journal metric approach. *Scientometrics*, 73, (1) , 53-78.
Nederhof, A. J., Zwaan, R. A., De Bruin, R. E., Dekker, P. J. (1989). Assessing the usefulness of bibliometric indicators for the humanities and the social sciences. *Scientometrics*, 15 : 423–435.
Nock, D. A. (2001). Careers in print: Canadian sociological books and their wider impact, 1975–1992. *Canadian Journal of Sociology/Cahiers canadiens de sociologie*, 26: 469–485.
Price, D. J. (1970). Citation measures of hard science, soft science, technology, and non-science. In: C. E. Nelson, D. Pollack (Eds), *Communication Among Scientists and Engineers*. Lexington, Mass., Lexington books.
Schloegl, C., & Gorraiz, J. (2010). Comparison of citation and usage indicators: The case of oncology journals. *Scientometrics*, 82(3), 567–580.
Schloegl, C., & Gorraiz, J. (2011). Global Usage Versus Global Citation Metrics: The Case of Pharmacology Journals. *Journal of the American Society for Information Science and Technology*, 62(1):161–170.
Shepherd, P.T. (2007). The feasibility of developing and implementing journal usage factors: a research project sponsored by UKSG. *Serials: The Journal for the Serials Community*, 20(2):117-123.
Small, H. G., Crane, D. (1979). Specialties and disciplines in science and social science: an examination of their structure using citation indexes. *Scientometrics*, 1 : 445–461.
Thompson, J. W. (2002). The death of the scholarly monograph in the humanities? Citation patterns in literary scholarship. *Libri*, 52 (3) : 121–136.
Wan, J.-K., Hua, P.-H., Rousseau, R., & Sun, X.-K. (2010). The journal download immediacy index (DII): Experiences using a Chinese full-text database. *Scientometrics*, 82(3), 555–566.

RELEVANCE AND FOCUS SHIFT: NEW METRICS FOR THE GRANT EVALUATION PROCESS PILOT TESTED ON NIH GRANT APPLICATIONS (RIP)

Duane E. Williams[1], Leo DiJoseph[1], James Corrigan[2], Elizabeth Hsu[2], Yvette R. Seger[1], Samantha Finstad[2], Emily J. Greenspan[3], Jerry S.H. Lee[3], Joshua D. Schnell[1]

[1] *duane.williams@thomsonreuters.com*
Thomson Reuters, Rockville, MD (USA)

[2] *corrigan@mail.nih.gov*
Office of Science Planning and Assessment, National Cancer Institute, National Institutes of Health Bethesda, MD (USA)

[3] *emily.greenspan@nih.gov*
Center for Strategic Scientific Initiatives, National Cancer Institute, National Institutes of Health Bethesda, MD (USA)

Abstract
Among the challenges faced by program staff in research funding organizations is obtaining an early assessment of the suitability of grant applications received in response to new funding announcements. Here we present two new metrics that use text mining to provide rapid and objective characterization of grant applications. This pilot study assesses the relevance and focus shift of grant applications submitted to the National Cancer Institute's (NCI) Provocative Questions (PQ) Initiative (RFA-CA-11-011 and RFA-CA-11-012). Relevance is measured by comparing the titles and abstracts of PQ grant applications to the background text on the PQ website summarizing the intent, goals and feasibility of each PQ. Focus Shift measures the similarity between PQ applications and prior applications submitted to the National Institutes of Health (NIH). Our results found the majority of applications to be relevant, but the relevance scores varied significantly by topic. Of the applications with very low focus shift scores, manual review of a subset found that 25-50% were very similar in scientific approach to previously submitted grant applications to other funding opportunities. The primary limitations of our automated approach are that similarity measurements are sensitive to the comparison text and often unable to distinguish subtle text differences.

Conference Topic
Science Policy and Research Evaluation: Quantitative and Qualitative Approaches (Topic 3).

Introduction
Automated assessments of grant portfolios may allow research funding agencies to incorporate additional objective and transparent metrics into funding decisions.

Here we apply automated text similarity calculations in novel ways to aid funding agencies in understanding their grant application portfolios. The use of text similarity algorithms is well established, but most commonly associated with the identification of plagiarized work (Errami, et al., 2010) (Bailey, 2002)(Long, 2009). In this study, we examine the first round of applications submitted to the National Cancer Institute's (NCI) Provocative Questions (PQ) Initiative (RFA-CA-11-011 and RFA-CA-11-012). This initiative sought to challenge the scientific community to creatively think about and answer important, but non-obvious or understudied questions in cancer research. To aid program staff in assessing the success of the initiative in attracting relevant and novel proposals, we use text similarity to assign numeric values for the relevance and focus shift of the grant applications. The relevance score is intended to assess how well the application responds to the core aspects of the PQ. Focus shift is intended to measure the extent to which an application represents a distinct approach from prior submitted applications. Both measurements are proxies for manual review aimed at providing program staff with a quick and objective overview of their grant applications. In this study, each application was assessed and assigned a single relevance score relative to the text used within the public description of the funding opportunity in the Request for Applications (RFA) documents. Two focus shift values were calculated for each PQ application. The first was in comparison to the investigator's own prior work ("by-self"), and the second was in comparison to NIH grants received from other investigators ("general"). The scheme for calculating these two values is presented below.

Methods

Calculating text distances

Text similarity scores were obtained using the FREETEXTTABLE function in Microsoft® SQL Server™, which is based on the Okapi BM25 algorithm.[139] Text similarity scores were converted to text distances in the range from 0 (similar) to 1 (dissimilar). The text distance scale from 0 to 1 is illustrated in **Figure 1**, along with a suggested interpretation of text distance as a measure of relevance and focus shift.

The text similarity scores are based on a Term Frequency/Inverse Document Frequency calculation (TF/IDF). As such, the values that result depend on the selection of pertinent documents that provide a corpus of text. *Relevance* measurements were made using a single document corpus of PQ applications and other similar grant applications identified as being coincidentally relevant to the PQ based on earlier unpublished work. *Focus shift* measurements were made using a single document corpus of the PQ applications and a comparison cohort of NIH applications received prior to the publication of the PQ RFAs. In calculating

[139] Microsoft Corporation (2008). SQL Server 2008 R2. How Search Query Results Are Ranked (Full-Text Search). Retrieved from http://msdn.microsoft.com/en-us/library/ms142524(v=sql.105).aspx.

focus shift for each PQ application, the comparison cohort was partitioned into two subsets based on whether the prior application was submitted by the same investigator (the "by-self" subset) or by different investigators (the "general" subset).

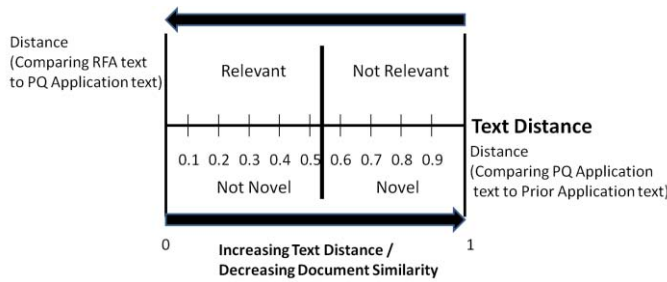


Figure 14. Text distance scale and suggested interpretation as relevance and focus shift. The vertical line at 0.53 denotes the threshold used for a binary classification of the PQ applications.

Text distances as scaled texts similarity scores

The call to the FREETEXTTABLE function returns a similarity score for each ordered pair of documents (Score(Left, Right)). The score has a documented absolute range of 0 to 1000, with the highest score corresponding to the most similar documents. Using the formula below, scores were converted to distance[140], so that a case with maximal similarity score is converted to a distance of 0, and a 0 similarity score is converted to a distance of 1 (the maximum distance). We experimented with several ways of defining what represented a maximal similarity score (e.g., whether to include a document self comparison or allow a hypothetical document comparison that would return 1000).

$$Distance\ (Left,Right) = \frac{max(all\ Scores\ from\ Left\ to\ any\ right\ side\ document) - Score(Left,Right)}{max(as\ above)}$$

We defined maximal similarity based on a manual review of a sample of scored document pairs. For relevance, this value was the highest score for all documents compared to the RFA. For the focus shift by-self measurement, the value was set to the highest score obtained in any of the by-self calculations. For the focus shift measurement relative to the general subset, we expanded the right side corpus to include the PQ applications, so that the maximal similarity was obtained from comparing a PQ application text to itself.

[140] This is not a distance in the strict sense of a mathematical metric or pseudo-metric. In particular, distance (left, right) is generally not equal to distance (right, left).

Relevance and focus shift measurement definitions and thresholds

After calculating the scaled text distances, we formally defined *relevance* as 1-relevance text distance, and *focus shift* (both forms) as the minimum of all focus shift distances measured for a given PQ application relative to either the by-self or general subsets of prior applications. For the thresholds, rather than attempting to calibrate the text distances (scaled similarity scores) against the results of expert manual comparison, we set the more modest goal of determining a fixed distance threshold value to classify applications as either relevant or not relevant, and as either focus shifted or not focus shifted. The thresholds were 0.53 for focus shift (both forms) and 0.47 for relevance.

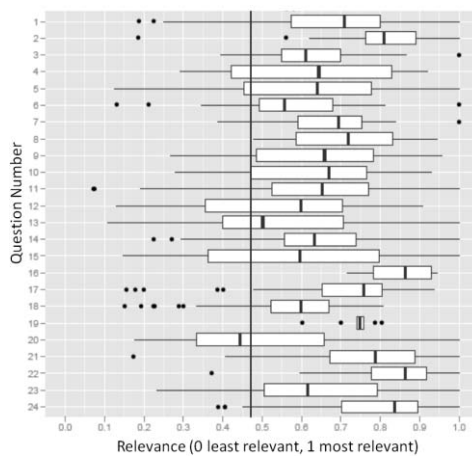


Figure 2. Relevance of PQ application text to RFA text

Results and Discussion

Relevance

Of the 754 PQ Applications, 614 (81.4%) were classified as relevant by score. Box plots of the measured relevance of all PQ applications are shown in **Figure 2** (Wickham, 2009) (R Development Core Team, 2012). The portion of the distribution to the ***right*** of 0.47 represents the applications that were classified as relevant to the background text on the PQ website for each question. The graph shows a high degree of variability in relevance among the applications for particular questions and significantly different distributions across the PQs. Results from a manual review suggest that our current approach for measuring relevance using text comparisons could be enhanced by a more sophisticated method that could appropriately account for semantic differences within the text. We are currently working to enhance our model to reflect these subtleties. Additionally, relevance scores were found to be highly influenced by the target text used. The results presented here were limited to the text used publicly to

describe the PQs.[141] However, this text often has specific references to therapeutic agents and methodologies designed to give examples of possible approaches to a respective PQ. When grant applications cite the same examples or even restate the problem, the scores could be inflated. We are currently working to address these potential pitfalls by preprocessing the target text used, which might include augmenting the text with other more generic descriptions of the problem. We should also note that automated approaches are generally limited in their ability to distinguish subtle differences in language that may indicate to human readers substantial differences in the content.

Focus shift

Since it is expected that investigators will tend to carry over ideas from prior research, achieving a focus shift classification in comparison to one's own prior applications – the by-self subset – was expected to pose a challenge. Conversely, it was expected that finding a very similar scientific approach within a general prior application (which excludes the investigator's own applications) would be less likely. Of the 754 PQ Applications, 39 (5.2%) were classified as focus shifted relative to the by-self prior subset score and 271 (35.9 %) were classified as focus shifted relative to the general subset score. Box plots of the by-self and general forms of the focus shift measurement for all PQ applications are shown in **Figures 3a** and **3b**, respectively. The portion of the distribution to the ***right*** of 0.53 represents those applications that were focus shifted relative to the prior applications for each question.

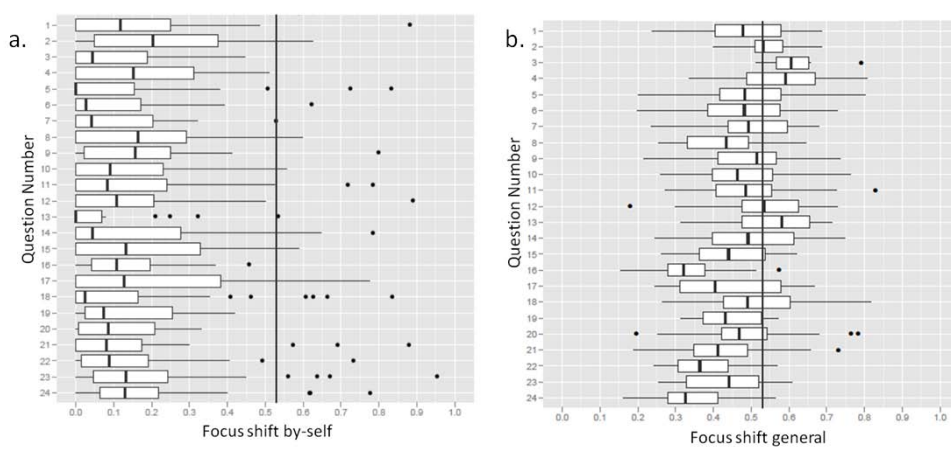


Figure 3. Box plots of PQ application focus shift relative to prior (a) by-self applications and (b) NIH general applications

[141] PQ RFA text used in this analysis can be found online at www.provocativequestions.nci.nih.gov.

Degree of Scientific Similarity

To better understand the correlation of our focus shift measurements with actual scientific similarity between two grant applications, we conducted a manual review using subject matter experts with a subset of grant applications with very low focus shift measures (minimum distance <0.05). Using this criterion, 41% (311/754) of applications to the PQs were identified as similar, 25% (189/754) of grants were similar to unfunded prior grants submitted by at least one of the principal investigators on the PQ application, and 12% (88/754) were similar to funded prior grants with publications.

The manual review was conducted on 40 applications subdivided into two groups based on the nature of the prior application to which they were most similar:

1. Applications similar to unfunded prior grants
2. Applications similar to funded prior grants with publications

This review found that PQ applications with low focus shift by-self measures cannot be assumed to have been reused from prior grants (see **Table 1**). Of the applications with low focus-shift relative to prior funded grants with publications, a larger percentage was found to be an extension of the prior work (45%) than the percentage with similar scientific approaches to prior grants (25%). PQ applications that were similar to prior unfunded grant applications had a greater likelihood of having similar scientific approaches to previously submitted grant applications (55%).

Table 10. Results from manual review of 40 PQ applications with very low novelty by-self distances (distances <0.05).

Classification	*Similar to unfunded grant applications*	*Similar to funded grant applications with Publications*
Similar scientific approach to prior application	55%	25%
Similar background/stage setting, scientific approach substantially different	30%	30%
Extensions of prior work	15%	45%

Possible extensions of this work may focus on identifying whether applications with similar scientific approaches were more successful in the review process than applications that represent a true focus shift. We will also investigate whether the trends in scientifically similar applications vary by specific PQ question to which the application responded. If there is a correlation between these applications and specific questions, it may inform the program staff regarding the accessibility of different questions to different audiences.

Relevance and focus shift quadrants

For the 702 PQ applications for which prior by-self applications were found, we now have 2 sets of paired values (focus shift by-self, relevance), and (focus shift

general, relevance). For the remaining 52 applications we have only the (focus shift general, relevance) pair. In this section, we examine the distribution of these paired values. As shown below in **Figure 4**, in Cartesian coordinates of the (focus shift, relevance) values, the 2 thresholds define 4 quadrants in which a given application can be found:

- Neither focus shifted nor relevant, lower left quadrant, abbreviated as **
- Focus shifted but not relevant, lower right quadrant, abbreviated as Fs*
- Relevant but not focus shifted, upper left quadrant, abbreviated as *R
- Focus shifted and relevant, upper right quadrant, abbreviated as FsR

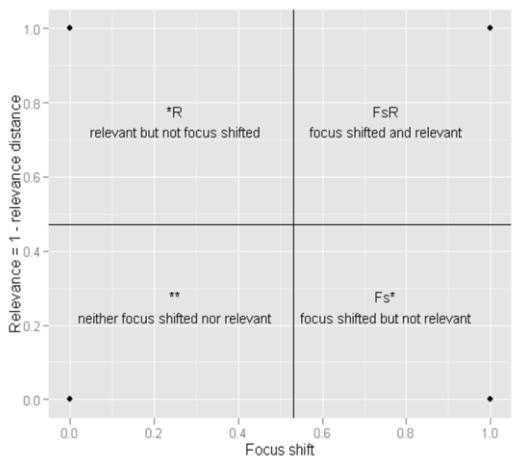


Figure 4. Focus shift /relevance quadrants.

Overall distributions of PQ applications over the FsR quadrants

Table 2 shows the overall distribution of PQ applications across the focus shift/relevance quadrants, using the focus shift by-self measurement. This table includes all 754 applications– those with no prior by-self applications were classified into either the (**) or (*R) quadrants depending on whether they were relevant. **Table 3** shows the quadrant distribution using the focus shift general measurement.

Table 2. Quadrant distribution using focus shift by-self.

Focus shift by-self and relevance classification	*Description*	*PQ application count*	*Percentage of applications*
FsR	focus shifted and relevant	26	3.4%
*R	relevant but not focus shifted	588	78.0%
Fs*	focus shifted but not relevant	13	1.7%
**	neither focus shifted nor relevant	127	16.8%

Table 3. Quadrant distribution using focus shift general.

Focus shift general and relevance classification	*Description*	*PQ application count*	*Percentage of applications*
FsR	focus shifted and relevant	182	24.1%
*R	relevant but not focus shifted	432	57.3%
Fs*	focus shifted but not relevant	89	11.8%
**	neither focus shifted nor relevant	51	6.8%

Conclusions

This work represents a first step toward the use of automated text mining algorithms to inform the grant evaluation process. The primary limitation of our current approach to calculate focus shift and relevance is that when two bodies of text are found to be similar, it may represent a similarity of background and stage setting rather than a similarity of the experimental approach. Generally, focus shift measurements were found to be more accurate than relevance in terms of their agreement with manual assessment of the scientific similarity between documents. Re-examining the choice of text used for analysis is likely to show promise in improving the confidence in the meaning of both focus shift and relevance scores; both measurements may be improved by the inclusion of the specific aims section of grant applications. Another important next step is to use a more sophisticated text mining approach that accounts for semantic relationships within the documents.

Bibliography

Bailey, B. (2002). Duplicate publication in the field of otolaryngology-head and neck. *Otolaryngol. Head Neck Surg.*, 211-216.

Errami, M., Sun, Z., George, A., Long, T., Skinner, M., Wren, J., et al. (2010). Identifying duplicate content using statistically improbable phrases. *Bioinformatics , 26* (11), 1453-1457.

Long, T. e. (2009). Responding to possible plagiarism. *Science , 323*, 1293-1294.

R Development Core Team. (2012). R: A Language and Environment for. Vienna: R Foundation for Statistical Computing.

Wickham, H. (2009). *ggplot2: elegant graphics for data analysis*. New York: Springer.

RELEVANCE DISTRIBUTIONS ACROSS BRADFORD ZONES: CAN BRADFORDIZING IMPROVE SEARCH?

Philipp Mayr[1]

[1] *philipp.mayr@gesis.org*
GESIS – Leibniz Institute for the Social Sciences, Unter Sachsenhausen 6-8, 50667 Cologne, Germany

Abstract
The purpose of this paper is to describe the evaluation of the effectiveness of the bibliometric technique Bradfordizing in an information retrieval (IR) scenario. Bradfordizing is used to re-rank topical document sets from conventional abstracting & indexing (A&I) databases into core and more peripheral document zones. Bradfordized lists of journal articles and monographs will be tested in a controlled scenario consisting of different A&I databases from social and political sciences, economics, psychology and medical science, 164 standardized IR topics and intellectual assessments of the listed documents. Does Bradfordizing improve the ratio of relevant documents in the first third (core) compared to the second and last third (zone 2 and zone 3, respectively)? The IR tests show that relevance distributions after re-ranking improve at a significant level if documents in the core are compared with documents in the succeeding zones. After Bradfordizing of document pools, the core has a significant better average precision than zone 2, zone 3 and baseline. This paper should be seen as an argument in favour of alternative non-textual (bibliometric) re-ranking methods which can be simply applied in text-based retrieval systems and in particular in A&I databases.

Introduction
The perceived expectations of users searching the web are that retrieval systems should list the most relevant or valuable documents in the result list first (so-called relevance ranking). More approaches appear that draw on advanced methods to produce relevant results and alternative views on document spaces. Google PageRank and its derivations (see e.g. Lin, 2008) or Google Scholar's citation count are just two popular examples for informetric-based rankings applied in Internet search engines.
Distributed search across multiple A&I databases will also generate large and heterogeneous document sets with the effect that users are confronted with a massive load of results from different scientific domains, even for specific research topics. Furthermore, empirical tests with typical A&I databases like Medline show that conventional term frequency - inverse document frequency (tf-idf) best match models and especially recent web-based ranking methods implemented in search engines (originally for web pages) are not always appropriate for search in heterogeneously collected scholarly metadata documents.

In this paper we want to apply and evaluate a non-textual ranking technique, called Bradfordizing. Introduced by H.D. White (1981), Bradfordizing is a bibliometric method to reorganize a search result for a topic. Bradfordizing is set up by applying the following procedure:

"… that is sorting hits (1) by the journal in which they appear, and then sorting these journals not alphabetically by title but (2) numerically, high to low, by number of hits each journal contains. In effect, this two-step sorting ranks the search output in the classic Bradford manner, so that the most productive, in terms of its yield of hits, is placed first; the second-most productive journal is second; and so on, down through the last rank of journals yielding only one hit apiece." (White, 1981: p. 47).

Bradford Law

Journals play an important role in the scientific communication process. They appear periodically, they are topically focused, they have established standards of quality control and often they are involved in the academic gratification system. Metrics like the famous impact factor are aggregated on the journal level. In some disciplines journals are the main place for a scientific community to communicate and discuss new research results. These examples shall illustrate the impact journals bear in the context of science models (Börner et al., 2011). Modeling science or understanding the functioning of science has a lot to do with journals and journal publication characteristics. These journal publication characteristics are the point where Bradford law can contribute to the larger topic of science models.

Bradford law of scattering bases on literature observations the librarian S. Bradford has been carried out in 1934. His findings and after that the formulation of the bibliometric model stand for the beginning of the modern documentation (Bradford, 1948) – a documentation which founds decisions on quantifiable measures and empirical analyses. The early empirical laws described by Lotka, Zipf and of course Bradford are landmark publications which still influence research in scientometrics (Bookstein, 1990), but also in other research communities like computer science or linguistics. In brief, scientometric and informetric research investigates the mathematical descriptions and models of regularities of all observable objects in the library and information science area. These objects include authors, publications, references, citations, all kinds of texts etc. Bradford's work bases on analyses with journal publications on different subjects in the sciences.

Fundamentally, Bradford law states that literature on any scientific field or subject-specific topic scatters in a typical way. A core or nucleus with the highest concentration of papers - normally situated in a set of few so-called core journals - is followed by zones with loose concentrations of paper frequencies (see Figure 1 for a typical Bradford distribution). The last zone covers the so-called periphery journals which are located in the model far distant from the core subject and normally contribute just one or two topically relevant papers in a defined period.

Bradford law as a general law in informetrics can successfully be applied to most scientific disciplines, and especially in multidisciplinary scenarios (Mayr, 2009). Bradford describes his model in the following:

"The whole range of periodicals thus acts as a family of successive generations of diminishing kinship, each generation being greater in number than the preceding, and each constituent of a generation inversely according to its degree of remoteness." (Bradford, 1934)

Bradford provides in his publications (1934, 1948) just a graphical and verbal explanation of his law. A mathematical formulation has been added later by early informetric researchers. Bradford`s original verbal formulation of his observation has been refined by Brookes (1977) to

$$G(r) = k \ln\left\{\frac{a+r}{a}\right\} \quad (1)$$

Where G(r) is the cumulative distribution function, k and a are constants, and r is the rank 1,2,…n.

The result of the application of this formula is often called a rank-order distribution of the items in the samples. In the literature we can find different names for this type of distribution, e.g. "long tail distribution", "extremely skewed", "law of the vital few" or "power law" which all show the same properties of a self-similar distribution.

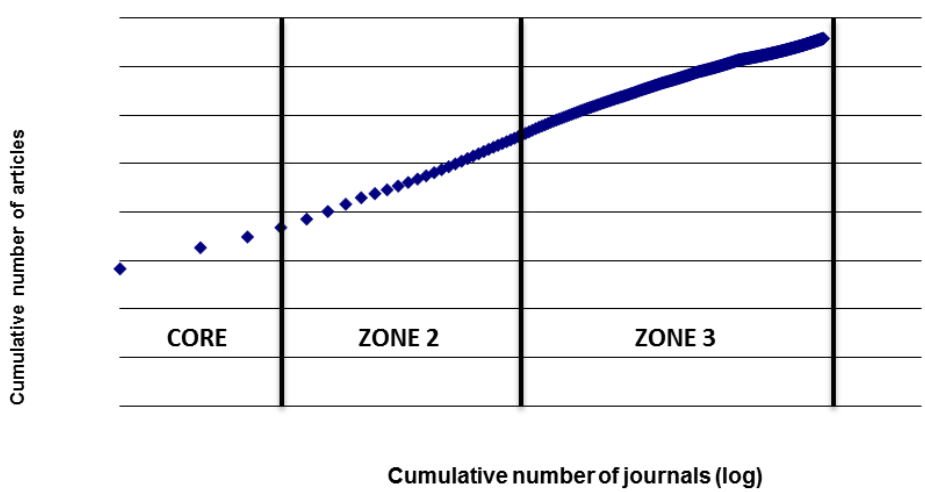


Figure 1. A typical Bradford distribution: Core, Zone 2 and Zone 3 (so-called periphery). The cumulative number of journals (x-axis) is displayed on a logarithmic scale.

In the past, Bradford law is often applied in bibliometric analyses of databases and collections e.g. as a tool for systematic collection management in library and information science. This has direct influence on later approaches in information science, namely the development of literature databases. The most common known resource which implements Bradford law is the Web of Science (WoS). WoS focuses very strictly on the core of international scientific journals and consequently neglects the majority of publications in successive zones.
To conclude this section, Bradford law is relevant for scholarly information systems due to its structuring ability and the possibility to reduce a large document set into a core and succeeding zones. As a consequence, modeling science into a core (producing something like coreness) and a periphery always runs the risk and critic of disregarding important developments outside the core.

Bradfordizing

Bradfordizing, originally described by White (1981), is a simple utilization of the Bradford law of scattering model which sorts/re-ranks a result set accordingly to the rank a journal gets in a Bradford distribution. The journals in a search result are ranked by the frequency of their listing in the result set (number of articles in a certain journal). If a search result is bradfordized, articles of core journals are ranked ahead of the journals which contain an average number (Zone 2) or only few articles (Zone 3) on a topic (compare the example in Figure 1). This re-ranking method is interesting because it is a robust and quick way of sorting the central publication sources for any query to the top positions of a result set.
Bradfordizing shows the following advantages: a) a structured view on a result set which is ordered by journals; b) an alternative view on publication sources in an information space which is intuitively closer at the research process than statistical methods (e.g. best match ranking) or traditional methods (e.g. exact match sorting); c) an approach to switch between the search modus e.g. starting with directed term searching and changing to a browsing mode (Bates, 2002) an improvement of relevance distribution between the journal zones, recently investigated (Mayr, 2009).
In principle, the ranking technique Bradfordizing can be applied to any search result with a minimum of 100 documents from one specific document type (e.g. journal articles). Generally Bradfordizing needs 100 or more documents because smaller document sets show too little scattering to divide the result into meaningful zones.
Bates' paper (2002) is interesting in our context because it brings together Bradford's Law (1934), information seeking behavior and IR (compare Wolfram, 2003, Garfield, 1996). Bates postulates "… the key point is that the distribution tells us that information is neither randomly scattered, nor handily concentrated in a single location. Instead, information scatters in a characteristic pattern, a pattern that should have obvious implications for how that information can most successfully and efficiently be sought."

The main task of this paper is to evaluate the effect when applying Bradfordizing to topical document sets from A&I databases. We want to answer the following question: Does Bradfordizing improve the ratio of relevant documents in the first third (core) compared to the second and last third (zone 2 and zone 3, respectively)?
The implementation of Bradfordizing in a typical digital library (DL) should be an alternative ranking option used to re-build and structure a result set. The intention is to list more relevant documents for a topic in the first third of a re-ranked result set. The re-ranking should be interpreted by users as a value-added due to the new structure and the relevance concentration of the listed documents after Bradfordizing. Furthermore Bradfordizing can be a helpful service to positively influence the search process. The opening up of new access paths and possibilities to explore document spaces for academic search questions can be a plausible value-added for users.
In the following section we will describe the research questions and methods used in our study (see Mayr, 2009).

Methods

In this paper we seek to answer the following research questions:

1. Is a re-ranking of documents according to Bradfordizing (ranking journal productivity or core journals first) a measurable added value for searchers?

The re-ranking of content to the most frequent sources (extracting the nucleus) can, for example, be a helpful access mechanism for browsing and initial search stages, especially for novice researchers in a discipline. Evaluation of the utility of such a simple re-ranking mechanism is still a desideratum.

2. Are the documents in the nucleus (core journals) of a bradfordized list more often relevant for a topic than items in succeeding zones with lower productivity?

Compared to traditional text-based ranking mechanisms, the bibliometric re-ranking technique Bradfordizing offers a completely new view on result sets, which have not been implemented and tested in heterogeneous database scenarios with multiple collections to date. This requires proving on a larger scale via intellectual assessments.

3. Can Bradfordizing be applied to document sources other than journal articles?

Few analyses show that monograph literature can be successfully bradfordized. But is this a utility for searchers? Other document types (proceedings, grey literature etc.) have to be equally proven.
In our study we focus on document sets from conventional subject-specific A&I databases. We have decided for a laboratory-based IR approach. Intellectual assessments of document relevance were performed following the classical IR evaluation experiments at TREC (e.g. Voorhees, 2007) and Cross-Language Evaluation Forum (CLEF). First of all, the organizers of a retrieval conference

like CLEF provide a test collection and a set of topics adequate to this test document corpus. Afterwards, participants apply their individual retrieval algorithms and systems while retrieving these topics (25 different topics each year in CLEF) in the test collection. Each participating retrieval system produces one or more ranked lists (called run) and sends these results back to the organizers. The organizers pool the documents from the retrieval runs for each topic and give the merged document pools away for objective intellectual relevance assessment. All documents in the document pools undergo binary assessment (relevant or irrelevant for a topic) by trained jurors (normally relevance is not binary (see Saracevic, 1975, Mizzaro, 1997 or White, 2007). The jurors perform the assessments on the basis of a short guideline.

We can hypothesize for our experiment: If the ratio of relevant documents, measured in precision (p), is the same in all three equally sized zones, then Bradfordizing has no effect on the distribution of relevant documents in the whole document pool. If the relevance ratio p in the first zone after re-ranking (core) is lower than p in the succeeding zones (zone 2 and zone 3), then Bradfordizing produced a falloff in precision. But if the ratio p of relevant documents in the core is higher than in other zones, and that is what we expect, then Bradfordizing improves the search result (measured in p) and consequently has a positive effect on search.

For this study, topics, documents and intellectual assessments from two evaluation initiatives have been analyzed: document pools from the GIRT-corpus in CLEF and the KoMoHe evaluation project (see Mayr & Petras, 2008). Our study analyzed scientific literature (journal articles and monographs) from social and political sciences, economics, psychology and medical science databases (see Table 1). Documents from the following database were included: SOLIS, SoLit, USB Köln Opac, World Affairs Online, Psyndex and Medline.

Table 1. Overview of the analyzed topics and documents in the IR experiments.

	CLEF	**KoMoHe**
Project period	2003-2007	2007
Number of topics	125	39
Domain, discipline	Social and political sciences	Social sciences, political sciences, economics, psychology and medical science
Assessed documents total	65,297	31,155
Journal articles bradfordized	18,112	17,432
Monographs bradfordized	11,045	4,900
Databases	2 (1)	6

We retrieved, analyzed and intellectually assessed 164 different standardized topics which yielded more than 96,000 documents from all the above domains. More than 51,000 assessed documents could be bradfordized.

The analysis of the data sets can be divided into three steps.

1. The document types journal articles and monographs are extracted from the document pool. Each document type and topic is analysed separately.
2. Each document set for a topic will be re-ranked according to Bradfordizing and divided into equally sized zones (core, z2 and z3).
3. The relevance assessments of the documents in the three zones are matched and aggregated zone by zone.

Average precision for each topic and zone can be calculated afterwards. We define the precision as the ratio of relevant documents out of all documents.
We calculate the average precision for each zone (core, zone 2 and zone 3) and baseline precision for the whole document pool (see Table 2 for an example).

Table 2. Example of the applied precision calculation for the CLEF-topic no. 171 "Computers in everyday life".

	Retrieved	Relevant	Precision
Core	73	41	0.56 (P core)
Zone 2	65	25	0.38 (P z2)
Zone 3	70	14	0.20 (P z3)
Total	208	80	0.38 (P baseline)

Results

The average precision for 164 tested topics from the projects CLEF and KoMoHe increases significantly after Bradfordizing (compare Table 3-6). So we can clearly verify research question 1. In this paper we show only precision values from analyses with journal articles. The largest precision benefit in both datasets is achieved between core and the last zone (zone 3). The improvements in Tables 4 and 6 marked with (*) are statistically significant based on the Wilcoxon signed-rank test and the paired T-Test. The improvements in the KoMoHe tests (see Tables 5, 6) are less significant, but average precision in the core outperforms precision in zone 3 impressively in all test series. Following this result we can clearly verify research question 2.

Table 3. Average precision for journal articles after re-ranking for five CLEF periods (N=125 topics). Core, Zone 2 (Z2), Zone 3 (Z3) and baseline.

CLEF articles	Topics	P core	P Z2	P Z3	P baseline
2003	25	0.294	0.218	0.157	0.221
2004	25	0.226	0.185	0.134	0.179
2005	25	0.310	0.240	0.174	0.239
2006	25	0.288	0.267	0.244	0.265
2007	25	0.278	0.256	0.217	0.248

Table 4. Average precision improvements for journal articles for five CLEF periods (N=125 topics). Core, Zone 2 (Z2), Zone 3 (Z3) and baseline.

CLEF articles	P@Core against P@Z3 in %	P@Core against P@Z2 in %	P@Z2 against P@Z3 in %	P@core against baseline in %
2003	86.56 (*)	34.57 (*)	38.63 (*)	32.65 (*)
2004	69.23 (*)	22.45	38.20	26.25 (*)
2005	78.03 (*)	29.05 (*)	37.95 (*)	29.52 (*)
2006	17.63	7.66	9.27	8.46
2007	28.18 (*)	8.31	18.35	11.77
Average 2003-2007	55.93 (*)	20.41 (*)	28.48 (*)	21.73 (*)

Table 5. Average precision for journal articles after re-ranking for three KoMoHe tests (N=39 topics). Core, Zone 2 (Z2), Zone 3 (Z3) and baseline.

KoMoHe articles	Topics	P core	P Z2	P Z3	P baseline
Test1	15	0.292	0.261	0.245	0.265
Test2	12	0.215	0.202	0.192	0.202
Test3	12	0.700	0.644	0.587	0.642

Table 6. Average precision improvements for journal articles for three KoMoHe tests (N=39 topics). Core, Zone 2 (Z2), Zone 3 (Z3) and baseline.

KoMoHe articles	P@Core against P@Z3 in %	P@Core against P@Z2 in %	P@Z2 against P@Z3 in %	P@Core against baseline in %
Test1	18.82	11.75	6.32	9.84
Test2	11.58	6.16	5.11	6.12
Test3	19.32 (*)	8.67 (*)	9.80 (*)	9.00 (*)
Average Test1-3	16.57 (*)	8.86	7.08 (*)	8.32 (*)

In general, the precision analyses with monographs in our tests show very similar results. The precision improvements after Bradfordizing (Bradfordizing of publishers) between zones are also positive but less significant than improvements with the journal articles (see research question 3).

Implementation

The proposed re-ranking service addresses the problem of oversized result sets by using the bibliometric method Bradfordizing. Bradfordizing re-ranks a result set of journal articles according to the frequency of journals in the result set such that articles of core journals are ranked ahead (see example in Figure 2). This re-ranking method is interesting for retrieval systems because it is a robust and quick way of sorting the central publication sources for any query to the top positions of a result set.

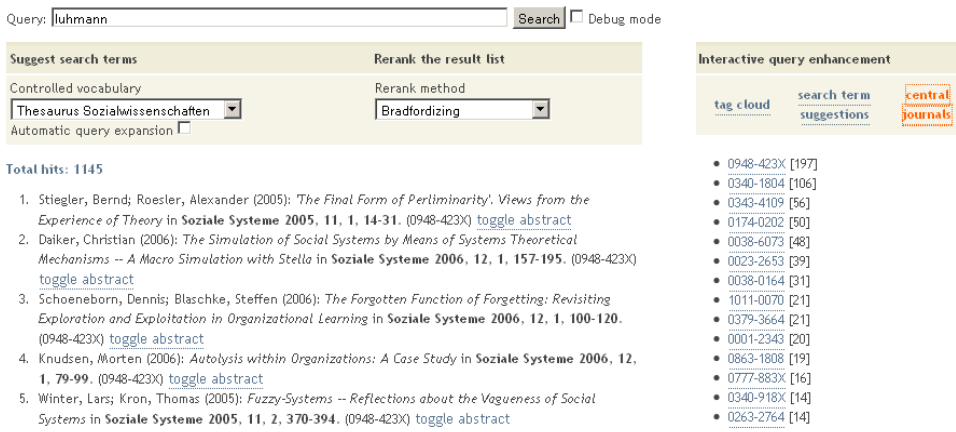


Figure 2. A bradfordized search for the search term "luhmann". ISSN numbers of journals and their productivity (article counts) are displayed on the left side of the screen. See research prototype under http://multiweb.gesis.org/irsa/IRMPrototype

The Bradfordizing procedure is implemented in the IRM prototype as a Solr plugin (see Figure 2 and a description of the prototype in Mayr et al., 2011). In a first step the search results are filtered with their ISSN numbers. The next step aggregates all results with an ISSN number. For this step we use a build-in functionality of our prototype engine Solr, the Solr faceting mechanism. Facets in Solr can be defined on any metadata field, in our case the "source" field of our databases. The journal with the highest ISSN count gets the top position in the result. The second journal gets the next position, and so on (see example in Figure 2). This procedure is an exact implementation of the original Bradfordizing approach. In the last step, the document ranking step, our current implementation works with a simple boosting mechanism. The frequency counts of the journals are used as boosting factors for documents in these journals. The numerical ranking value from the original tf-idf ranking of each document is multiplied with the frequency count of the journal (see Schaer, 2011). The result of this multiplication will be taken as ranking value for the final document ranking.

In principle, this ranking technique can be applied to any search result providing qualitative metadata (e.g. journal articles in literature databases). Generally, Bradfordizing needs 100 or more documents because smaller document sets often show too little scattering to divide the result into meaningful zones. Bradfordizing can be applied to document types other than journal article, e.g. monographs (cf. Worthen, 1975; Mayr, 2008, 2009). Monographs e.g. provide ISBN numbers which are also good identifiers for the Bradfordizing analysis.

To conclude, our implementation of re-ranking by Bradfordizing is a simple approach which is generic, adaptable to various document types and quickly implementable with build-in functionality. The only precondition for the application is the existence of qualitative metadata to assure precise identification and access to the documents. An evaluation of the value-added services of

Bradfordizing and other approaches has been published recently by Mutschke et al. (2011).

Discussion

The discussion of the re-ranking method Bradfordizing will focus on possible added-values and the positive and negative effects of this method. Some added-values appear very clearly. On an abstract level, re-ranking by Bradfordizing can be used as a compensation mechanism for enlarged search spaces with interdisciplinary document sets. Bradfordizing can be used in favor of its structuring and filtering facility. Our analyses show that the hierarchy of the result set after Bradfordizing is a completely different one compared to the original ranking. The user gets a new result cutout with other relevant documents which are not listed in the first section (in our experiment the top 10 documents) of the original list. Furthermore, Bradfordizing can be a helpful information service to positively influence the search process, especially for searchers who are new on a research topic and don't know the main publication sources in a research field. The opening up of new access paths and possibilities to explore document spaces can be a very valuable facility. Additionally, re-ranking via bradfordized documents sets offer an opportunity to switch between term-based search and the search mode browsing. It is clear that the approach will be provided as an alternative ranking option, as one additional way or stratagem to access topical documents (cf. Bates, 2002).

Interesting in this context is a statement by Bradford where he explains the utility of the typical three zones. The core and zone 2 journals are in his words "obviously and a priori relevant to the subjects", whereas the last zone (zone 3) is a very "mixed" zone, with some relevant journals, but also journals of "very general scope" (Bradford, 1934). Pontigo and Lancaster (1986) come to a slightly different conclusion of their qualitative study. They investigated that experts on a topic always find a certain significant amount of relevant items in the last zone. This is in agreement with quantitative analyses of relevance assessments in the Bradford zones (Mayr, 2009). The study shows that the last zone covers significantly less often relevant documents than the core or zone 2. The highest precision can very constantly be found in the core.

To conclude, modeling science into a core and a periphery – the Bradford approach – always runs the risk and critic of disregarding important developments outside the core. Hjorland and Nicolaisen (2005) recently started a first exploration of possible side effects and biases of the Bradford methods. They criticized that Bradfordizing favours majority views and mainstream journals and ignores minority standpoints. This is a serious argument, because by definition, journals which publish few papers on specific topics have very little chance to get into the core of a more general topic. A counter-argument could be that the Bradfordizing approach is just an application which is working on existing document sets. The real problem is situated before, in the development of a data set, especially in the policy of a database producer.

Conclusions

An evaluation of the method and its effects was carried out in two laboratory-based information retrieval experiments (CLEF and KoMoHe) using a controlled document corpus and human relevance assessments (see Ingwersen & Järvelin, 2005 for pros and cons of this methodology). The results show that Bradfordizing is a very robust and promising method for re-ranking the main document types (journal articles and monographs) in today's digital libraries (DL). The IR tests show that relevance distributions after re-ranking improve at a significant level if articles in the core are compared with articles in the succeeding zones. The items in the core are significantly more often assessed as relevant, than are items in zone 2 or zone 3. The largest increase in precision can typically be observed between core and zone 3. This has been called the Bradfordizing effect.

The results of our study can also be seen as a coalescence of Bradford Law in so far as Bradford did not postulate or observe a relevance advantage in the core. In Bradford's eyes all documents in his bibliographies were "relevant to a subject". His focus was the scattering of documents across journals, not the relevance distribution between document zones. According to Saracevic (1975), Bradford (1934) was one of the first persons to use the term relevant in our context ("relevant to a subject"). The results in this study show that articles in core journals are valued more often relevant than articles in succeeding zones (compare Garfield, 1996). This is an extension to the original conception of relevance distribution in the zones by Bradford. As we can empirically see, bibliometric distributions like Bradford distributions can also be described as "relevance related distributions" (Saracevic, 1975). The examination of relevance concentrations in our test series (CLEF and KoMoHe) show that there is not a massive concentration of relevant articles in the core, rather it is more a continuously decreasing of average precision from core to zone 3.

The relevance advantage in the core can probably be explained in that a) core journals publish more state-of-the-art articles, b) core journals are more often reviewed by peers in a certain field and c) core journals cover more aspects of the searched topic than journals in the peripheral zones. Further research is needed to clarify these questions.

Further research

After evaluating the positive relevance effect of Bradfordizing, our next goal is to go automatically from directed searching into a browsing mode. Starting with a subject-specific descriptor search, we will treat the query with our heterogeneity modules (Mayr & Petras, 2008) to transfer descriptor terms into a multi-database scenario. In a second step, the result lists from the distributed databases are combined, merged and re-ranked by users e.g. according to Bradfordizing. Step 3 could be the extraction of a result set of documents in the Bradford nucleus which can be delivered for browsing or other search stratagems. This browsing modus, based on automatically bradfordized lists, can be compared to the search technique which Bates terms "journal run."

The exploration of possible side effects and bias (see e.g. Nicolaisen & Hjorland, 2007) of this promising re-ranking method will be a next step. Recently Nicolaisen & Hjorland have criticized Bradfordizing: "Bradford analyses function discriminatorily against minority views … Bradford analysis can no longer be regarded as an objective and neutral method." This has to be proven on a larger empirical basis.

A comparison with other ranking and re-ranking methods would be highly desired. Techniques like bibliometric re-ranking (e.g. Bradfordizing described in this paper) or the application of social-network analysis techniques (e.g. co-authorship relationships in Mutschke, 2003) or other combinations of value-added services can and should be applied in digital libraries (DL) to improve IR (White 2005, 2007). Further research will focus on the implementation and evaluation of the method in a live system with different modules for improving retrieval (see Mutschke et al, 2011).

Acknowledgement

The KoMoHe project at GESIS ("Competence Center Modeling and Treatment of Semantic Heterogeneity") was the starting point and background of this study. The KoMoHe project was funded by the German Federal Ministry for Education and Research (BMBF) grant number 523-40001-01C5953. The retrieval prototype has been developed in the DFG project "Value-Added Services for Information Retrieval" (IRM) under project number: INST 658/6-1.

References

Bates, M. J. (2002). Speculations on Browsing, Directed Searching, and Linking in Relation to the Bradford Distribution. Paper presented at the Fourth International Conference on Conceptions of Library and Information Science (CoLIS 4).

Börner, K., Glänzel, W., Scharnhorst, A., & van den Besselaar. P. (2011). Modeling science: Studying the structure and dynamics of science." Scientometrics 89, 347–348.

Bradford, S. C. (1934). Sources of information on specific subjects. Engineering, 137(3550), 85-86.

Garfield, E. (1996). The Significant Scientific Literature Appears In A Small Core Of Journals. The Scientist, 10(17), 13.

Ingwersen, P., & Järvelin, K. (2005). The Turn - Integration of Information Seeking and Retrieval in Context. Dordrecht: Springer.

Lin, J. (2008). PageRank without hyperlinks: Reranking with PubMed related article networks for biomedical text retrieval. BMC Bioinformatics, 9.

Mayr, P. (2009). Re-Ranking auf Basis von Bradfordizing für die verteilte Suche in Digitalen Bibliotheken. Humboldt-Universität zu Berlin, Dissertation.

Mayr, P., Mutschke, P., & Petras, V. (2008). Reducing semantic complexity in distributed digital libraries: Treatment of term vagueness and document re-ranking. Library Review, 57(3), 213-224.

Mayr, P., Mutschke, P., Petras, V., Schaer, P., & Sure, Y. (2011). Applying Science Models for Search. In J. Griesbaum, T. Mandl, & C. Wormser-Hacker (Eds.), 12. Internationales Symposium für Informationswissenschaft (ISI 2011) (pp. 184–196). Hildesheim: vwh Verlag Werner Hülsbusch.

Mayr, P., & Petras, V. (2008). Cross-concordances: terminology mapping and its effectiveness for information retrieval. In 74th IFLA World Library and Information Congress. Québec, Canada.

Mizzaro, S. (1997). Relevance: The whole history. Journal of the American Society for Information Science, 48(9), 810-832.

Mutschke, P., Mayr, P., Schaer, P., & Sure, Y. (2011). Science models as value-added services for scholarly information systems. Scientometrics, 89(1), 349–364. doi:10.1007/s11192-011-0430-x

Mutschke, P. (2003). Mining Networks and Central Entities in Digital Libraries: a Graph Theoretic Approach Applied to Co-Author Networks. Paper presented at the Advances in Intelligent Data Analysis 5. Proceedings of the 5th International Symposium on Intelligent Data Analysis (IDA 2003), Berlin.

Nicolaisen, J., & Hjørland, B. (2007). Practical potentials of Bradford's law: A critical examination of the received view. Journal of Documentation, 63.

Saracevic, T. (1975). Relevance: A review of and a framework for the thinking on the notion in information science. Journal of the American Society for Information Science, 26(6), 321-343.

Schaer, P. (2011). Using Lotkaian informetrics for ranking in digital libraries. In C. Hoare & A. O'Riordan (Eds.), Proceedings of the ASIS&T European Workshop 2011 (AEW 2011). Cork, Ireland: ASIS&T.

Voorhees, E. M. (2007). TREC: Continuing information retrieval's tradition of experimentation. Communications of the ACM, 50(11), 51-54.

White, H. D. (1981). 'Bradfordizing' search output: how it would help online users. Online Review, 5(1), 47-54.

White, H. D. (2005). On Extending Informetrics: An Opinion Paper. In Proceedings of the 10th International Conference of the International Society for Scientometrics and Informetrics (pp. 442-449). Stockholm, Sweden: Karolinska University Press.

White, H. D. (2007). Combining bibliometrics, information retrieval, and relevance theory, Part 1: First examples of a synthesis. JASIST, 58(4), 536-559.

Wolfram, D. (2003). Applied informetrics for information retrieval research. Westport, CT: Libraries Unlimited.

RESEARCH COLLABORATION AND PRODUCTION OF EXCELLENCE: FINLAND 1995-2009

Hannes Toivanen[1] and Arho Suominen[2]

[1] *hannes.toivanen@vtt.fi*
VTT Technical Research Centre of Finland, Innovation and Knowledge Economy, P.O.Box 1000, 02044 Espoo, Finland

[2] *arho.suominen@vtt.fi*
VTT Technical Research Centre of Finland, Innovation and Knowledge Economy, Itäinen Pitkäkatu 4, Turku, P.O. Box 106, 20521 Turku, Finland

Abstract
This study uses complete-normalized counting in assessing credit for authorship and citations received, and argues that conventional bibliometric assessments used for policy development lead to misguided conclusions about how best research is created, and what type policies may promote research excellence. Exploring Finnish research 1995-2009 based on ISI data, we demonstrate that the nature of the Finnish "hot papers" (papers that receive most citations within two years after publication) doesn't correspond with the idealized vision of "high quality research" by being highly national and created by relatively small author teams. As such, it also resembles closely research with no impact, i.e. the non-cited papers. These two differ from the "other cited papers", which are authored by larger and highly international teams. While we describe the author team structure and national nature for different cohorts of scientific excellence, our central result is the observation that in terms production of excellence, whole citations created per author, small Finnish author teams are slightly more productive than large international author teams. We discuss at some length the methodological and policy implications of our results, especially as far as they give rise to the suspicion that conventional (Finnish) policy efforts to foster research excellence target the middle-tier papers and target poorly the best papers that resemble closely the worst ones. We also demonstrate how results and conclusions are highly dependent whether research excellence assessment focuses on papers or alternatively researchers. Finally, we consider how "scientific excellence" should be defined and measured in national contexts.

Conference Topic
Collaboration Studies and Network Analysis (Topic 6); Research Fronts and Emerging Issues (Topic 4); Science Policy and Research Evaluation: Quantitative and Qualitative Approaches (Topic 3).

Introduction
What type of research collaboration fosters excellence in knowledge creation? This remains a central challenge for sociology of science as well as research

policy. By using the complete-normalized counting in assessing credit for authorship and excellence with Finnish data for 1995-2009, this paper questions if conventional bibliometric assessments lead to misguided conclusions about what kind of research collaboration and what type policies may promote research excellence within national context.

In this context, we explore the relationship between research collaboration and scientific excellence. Our approach is based on differentiation three layers of science in Finnish research: The national *"hot papers"* consisting of the most visible and high-impact research, whereas the two other sets consist of other cited research and research without any citations. Methodologically, we demonstrate that bibliometric measures of research collaboration that do not recognize at article level the geographic locations of co-authorship (fractional counting), or relate papers' received citations to number of authors and their geographic location, are bound to provide incorrect assessments of the impact of research collaboration on excellence of science. Consequently, much of research indicators focus on papers, instead of actual researchers and research communities, undermining the credibility and effectiveness of policy incentives for research excellence. We use a method to credit citations received based on the share of institutional authorship, and demonstrate that results derived with such a method provide accuracy new frame for the assessment of the impact of research collaboration on research excellence.

This paper demonstrates the amount of "noise" included in conventional (uncalibrated) bibliometric measures by contrasting at article level the international and national perspective, count of citations per author, and geographic location of authorship. Besides of providing accurate and more realistic description of co-authorship, our basic policy concern is whether policy efforts to boost excellence through internationalism are really viable in the context of recent doubts on existing funding award criteria (Nicholson & Ioannidis, 2012).

Our results suggest that highly domestic and relatively small collaborative teams are most effective in producing of high-impact and most visible (as measured through citations) science in Finland. Large international author teams produce in absolute terms more citations, but Finnish participation in such teams is relatively marginal, weakening their impact on Finnish science community. Furthermore, smaller teams have more efficient ratio of received citations per author. Finally, we demonstrate that the Finnish "hot papers" science, (derived with our methods) resembles in terms of co-authorship structure that part of Finnish science that doesn't receive citations at all, raising questions about how to focus research policies fostering excellence.

Research collaboration, productivity, and excellence

One of the major changes in the university research environment is the introduction of a plethora of performance-based research funding systems. These systems are a research policy tool often directed implicitly or explicitly towards creating research excellence. Reviewed in detail by Hicks (2012), the various methods of using performance-based measures are in many countries indicator based, measuring the number of publications or publications and citations. The existing policy regime assumes that the international collaboration is a route towards increased citation and publication counts, thus to indicator based research excellence.

Although there can be little doubt that research collaboration is today the norm in practically all fields of science (Beaver, 2001), our understanding of the dynamics and impacts of research collaboration are much more murky. Indeed, Bozeman et al. (2012) note in their recent extensive review of research collaboration research that the existing literature doesn't really answer whether research collaboration really pays off, or under what conditions it would do so. In parts this might be due to a lack of valid indicators that would enable policy decisions (Edler & Flanagan, 2011)

Clearly one problem is the choice of right perspective to assess the productivity of research collaboration. Although on a global scale teams increasingly author most and the best papers (Jones, Wuchty, & Uzzi, 2008), there is increasingly literature that demonstrates that research collaboration is actually highly nuanced and differs by the field of inquiry, as well as by the size or level of development of country.

Indeed, the existing literature about the benefits of research collaboration is inconclusive. A series of studies have demonstrated the benefits of international collaboration. Looking at UK science, Katz and Hicks (1997) demonstrated that international co-authorship increased received citations much more than domestic collaboration. In case of chemistry, Glänzel and Schubert (2001) showed that whereas international co-authorship produced higher citation rates than purely domestic papers, it didn't contribute to the citation eminence. Also, the intensity of participation in research collaboration enhances quality. (Liao, 2011)

In contrast, a number of studies have cast doubts whether research collaboration really is an efficient way of fostering excellence. Ionnadis (2008) argued that fractional count of citations received reveals that large author team papers are less efficient in attracting citations than smaller teams. Lee and Bozeman (2005) demonstrated that collaboration is not a good predictor of publishing productivity or high citation rates, and identified the nature of the collaboration relationship as more significant factor than the author team size.

Reflecting the complexity of research collaboration as a phenomenon, Schmoch and Schubert (2008a) speculated that the increasing specialization in science could cause internationally co-authored papers to receive less citations: If increased specialization decreases the size of scientific fields, then the increasing absence of potential domestic collaboration partners would push scientists increasingly to international collaboration, and the overall shrinking of the field would result in smaller citation rates. Other identified drivers of international collaboration are the increasingly important role of instruments for knowledge production and access to high-cost research equipment. Finally, funding has been identified as one of the key drivers of international research collaboration, not least because science policies often assume that this is key instrument to improve the quality of science. (Defazio, Lockett, & Wright, 2009; Fraunhofer ISI, Idea Consult, & SPRU, 2009)

Although scientometric literature and evidence maintains that the relationships between research collaboration, productivity, and excellence is an fragmented and complex phenomenon, there are plenty of examples of national policy makers making blanket assumptions that international research collaboration fosters research excellence. (e.g. in Finnish context: Muhonen, Leino, & Puuska, 2012; Treudthardt & Nuutinen, 2012)

Authorship and citation impact

Although there certainly is some kind of relationship between research collaboration and citation impact, this "is a complex phenomenon that does not fit simple patterns," as Schmoch and Schubert (2008b) have noted. Reviewing methods to measure research collaboration and its impacts, Persson et al. (2004) argued that the uncritical use of basic indicators could easily lead to wrong conclusions, and recommended the use of normalization measures and relative indicators to strengthen validity of conclusions.

Here we focus on two well-known challenges of using citations to identify high-impact papers or research fields. The first one involves how to credit authorship and citations correctly for co-authored papers, and whether whole counting or fractional counting would provide significantly different results. (Gauffriau, Larsen, Maye, Roulin-Perriard, & Von Ins, 2007; Gauffriau, Larsen, Maye, Roulin-Perriard, & von Ins, 2008; Huang, Lin, & Chen, 2011; Wagner & Leydesdorff, 2005, 2012). This problem basically requires one to choose to focus on high impact papers or authors. The issue of fractional counting is repeated if one wants to analyze impacts in national contexts, which requires that credit for citations is allocated according to the geographical location of authors. (Toivanen, 2012a.)

The second challenge emerges from the highly differing citation traditions (i.e. frequencies) of different fields (Leydesdorff & Opthof, 2010; Waltman, van Eck,

van Leeuwen, Visser, & van Raan, 2011). If one employs citations as criteria to identify high-impact papers or authors, or successful fields, this easily skews selection if not controlled.

Research frontier

The notion of research frontiers implies the most advanced, path-breaking new scientific knowledge, and its measurement and identification has been a key theme of scientometrics (Garfield & Merton, 1979; Garfield & Small, 1989; Garfield, 2006). "Hot" science and technology research are research frontiers that have immediate, or concentrated, high impact on other research or technology. "Hot research" is typically defined as papers that receive the largest share of citations within certain time frame. Garfield and Small (1989) identified as "hot fields" clusters of papers that generated most citations within three years from publication.

There is relatively little research on analysing research frontiers in national contexts (Toivanen, 2012a), it's identification and measurement in such a context suffers from several of the difficulties associated with the assessment of research collaboration. Here we imply with research frontier the best and most visible research when measured as citations received. Towards this end, we use two methods to calculate received citations, explained in detail below. The point of separating Finnish research frontier from the rest of the research is to allow us to contrast the nature, structure, and dynamics of research collaboration among different research paper populations ranked among citation based "excellence".

Data and Methods

The study focuses on the research frontier, by limiting the data used to a relatively short citation windows. Excellence is looked through a proxy of citations accumulated in the selected narrow frontier citation window. Citations credited to authors based on a fractional (complete-normalized) counting scheme. The methodological selections are described in detail the following sub-sections.

Data

We use Web of Science data articles including at least one Finnish affiliated author from 1995 to 2009, provided in XML format on article level by Thomson Reuters in August 2012, including all WOS indexed fields. The data was processed with Vantagepoint software. For this research, we have restricted data to types Article, Proceedings paper, Meeting abstracts, and Reviews, totaling 141 554 papers. The data appears uniform for all other years except 2009, which has a drop in the share of Proceedings papers, lowering significantly the number of received citations and skewing the overall composition of document types. We look forward to fix this by re-calculating 2009 data with new data later.

For each year, we have also received all papers awarding citations to Finnish publications (totaling approximately two million records). Using this latter set and unique article level identifier, we estimated how many times each Finnish publication, published between 1995-2009, was cited during a 4-year citation window stretching from one year before the publishing year to two full years after. Naturally, we are able to record only WOS indexed citations.

Counting methods

In a perfect situation, the credit from a publication or its received citations would be accredited based on the effort the authors or resources an organization has put into the work. Bibliometrics is, however, limited by the meta information available in the bibliometric databases. This does not contain any knowledge on the actual work carried out by the authors or the nature of the authors' affiliations. With the limited data available, scholars should focus on understanding the limitations of different methods and describing explicitly the methodological selection made, thus creating reproducible results of which the limitations are clearly described.

Evaluating bibliometric results is dependent on the quantities and scoring methods used to calculate the results. Not to go as far as arguing that the varying use of terminologies and methods have created a crisis (Glänzel & Schoepflin, 1994) but the lack of consistency in describing results has significantly limited the reproducibility of results. Illustrated by an practical example by Chao et al. (2007), Seymour et al. (2007) and Liu et al. (2012) and the discussion that followed (Ho, 2009, 2012; P. Larsen, 2008; Seymour, 2008), being able to use a common shared vocabulary in describing the methodological options used to gain results is an credibility issue. This does not only challenge authors less familiar with bibliometrics who try to apply it, but also to scholars active in the field of scientometrics as pointed out by Larsen (2008).

In the case of directly policy focused bibliometric studies, credit for authorship is frequently obfuscated, leading to wrong-headed policy conclusions, priorities, strategies and whole-sale perceptions about the state of research. Reports published (Karlsson & Persson, 2012; Opetus- ja kulttuuriministeriö, 2011a, 2011b) often describe in lose detail the methodological options made and discuss little or none the limitations set by methodological decision made during the process – even though the underlying methodological selections would be sound. This sets significant practical implications if research funding is interconnected with studies where possible bias created through methodologies.

Discussion on the standardization of scientometric 'methods' was taken up by several researcher in a workshop in 1995, discussion subsequently published in 1996 (Scientometrics Vol 35 Issue 2). Much of the more recent literature on bibliometric counting methods (Gauffriau et al., 2007, 2008; Vinkler, 2001) is

drawn from the work published in 1996 (Bourke & Butler, 1996; McGrath, 1996; Vinkler, 1996). Recent efforts on creating a systematic approach on the measures of scientific publications counting was done by Gauffriau et.al (Gauffriau et al., 2007, 2008). The authors took a set theory approach to creating the foundation for standardized measures in scientific research by consistently defining four counting methods; complete counting (C), complete-normalized counting (CN), straight counting (S), whole counting (W) and whole-normalized counting (WN). In addition, Gauffriau et.al (2007, 2008) defined two notations basic unit of analysis (B) and object of study (O), which with the counting method enables an author to explicitly define the method of attributing a score in a bibliometric study.

Gauffriau et.al (2007, 2008) define three values of B and O, author, country, and institution, which should be selected for a study independently depending on the research objective. This selection should be complemented with an understanding of the limitations of the selected counting method. The selection of the previously mentioned counting method, unit of analysis and object of study create context and limitations for the results of the calculation made.

Looking at the counting methods in more detail, complete counting attributes each unit of analysis one credit. If for example setting the object of the study and basic unit to countries an article with two affiliations from Finland, one from Sweden and one from Denmark, Finland would receive two credits and Sweden and Denmark both one. In this, it is important to note the notation credit, to be distinguished from counting publications. In Huang et al. (2011)the authors do not consider complete counting as a "…reasonable approach" as they claim that the previous example would resulted in Finland producing two publications. This is of course not the case, and it is important note the difference between attributing credit and the number of publications. In CN counting all of the basic units share one credit based on the number of times each basic unit of analysis is mentioned in the publication. This method is also referred to as fractional counting (for example Aksnes, Schneider, & Gunnarsson, 2012)

In straight counting the first basic unit of analysis mentioned receives full credit of the publication that is one credit. First basic unit thus often refers to the first affiliation mentioned in the paper or the country of that affiliation. Studies have argued, although against logic, that straight counting results would correlate strongly with the results of other counting measures (see Lange, 2001). This has later been questioned by Zhao (2006), who discussed in detail the differences and possible biases of the straight counting approach.

In literature whole counting (W) has been described by several different names, such as full counting, normal counting and integer counting. The before mentioned describe a crediting method where if a basic unit (B) appears more

than one time it will still receive only one credit. Whole counting is often used when calculating citation indexes, and Moed (2005) argues that whole counting is a valid measure of participation. With whole-normalized counting (WN) each unique basic unit will share one credit, thus serving as a means of normalization. The above mentioned methodological option, described in detail by Gauffriau et al. (2007, 2008), are summarised in Table 1.

Table 1. Scores for different counting methods for a publication with one author from Denmark, two authors from Finland, and on author from Sweden, where one of the Finnish authors has the first affiliation mentioned in the document. (for details refer to Gauffriau et al., 2007, 2008)

	Complete counting C	Complete-normalized counting CN	Straight counting S	Whole counting W	Whole-normalized counting WN
Denmark	1	1/4	0	1	1/3
Finland	2	2/4	1	1	1/3
Sweden	1	1/4	0	1	1/3
Sum	4	1	1	3	1

Altogether, we should notice that often the selection of a method is not a question of right or wrong, but more a practical question of does the indicator measure what we think it measures? Using similar terminology and focusing the discussion on what we expect that a counting method actually measures would be valuable. For example, when looking at a smaller portion of the whole population of publications, such as focusing on research excellence and thus looking at a perceived top segment of the publications, we should be aware that the known limitations of bibliometric indicators are amplified when focusing on a smaller sample. In addition, in the case of research excellence we are forced to question if bibliometric measures have the ability to distinguish excellence from "very good" - as excellence might be defined by a number of capacities falling outside the grasp of bibliometric indicators (Rons & Amez, 2009).

The complexity of assigning credit is multiplied with accrediting citations to different basic units. (Gauffriau et al., 2007) Logically, when moving from assigning credit from one publication to dividing citation credits we increase the complexity of the problem. When assigning publication credits, the differences between counting methods are controlled by the relatively low deviation and mean value of an average number of basic units in a publication. With citations we are faced with a larger variation from publication to publication. This ultimately changes the dynamics of giving credit.

Citation counting methods are challenged by the difference in the patterns and traditions of publishing in different fields of science - resulting in a long-standing

debate on the use of normalization methods, such as a crown indicator (Moed, De Bruin, & Van Leeuwen, 1995) or normalisation at a publication level (Leydesdorff & Opthof, 2010), and the use of either whole counting or complete-normalized counting for country level research impact (Aksnes et al., 2012).

Counting methods applied in the study

The counting method applied in the study is complete-normalized counting. By using the research address field as a proxy for how work has been divided at a country level, we calculated for each article the share of domestic and foreign *institutional authors*. This allowed us to calculate Finnish institutional author share, i.e. complete-normalized count, (hereafter *FI Auth*) of all institutional authorship. More importantly, the "*FI Auth share*" share of citations received for each article (hereafter *FI FRC Citations*), measuring the what share of citations received can be credited to Finnish institutional authors. Taking the example from Table 1, the *FI Auth* is 2/4, and assuming that the article has received 12 citations the *FI FRC Citations* is 6.

For each year, we created two data sets according to total number of citations received during the citation window used. First data set is based on the whole, absolute, citations received (*ABS Citations*). Second data set is based on using the *FI FRC Citations* as a classifying factor. We have divided these two sets of annual publications 1995-2009 in to three echelons of research: First, the best and most visible research, or "*Hot papers*" or the "research front", defined here as the most cited 10% of publications that receive citations within the citation window. Second, the "*Other cited papers*" that receive citations less than the "*Hot papers*" within citation window, and, third, the "*Non-cited papers*".

The "*Hot papers*" consists of the roughly 10% of most cited papers among all papers that have received citations during the citation window. If it was not possible to apply the threshold directly, we included in "*Hot papers*" all papers receiving exactly the same number of citations as the paper at 10% threshold or nearest it. The "*Other cited papers*" (both "*FI FRC Citations"* and "*ABS Citations*") versions for each year) consist of all the papers that have received citations and that fell below the "*Hot papers*" citation threshold. "*Non-cited papers*" consists of annual sets of papers that have not received single citations, and is – obviously- the same for "*FI FRC Citations"* and "*ABS Citations*" data.

The approach used is based on several assumptions and limitations. First, citations are used as a measure of excellence. This is of course limited, as research excellence, not to mention personal excellence, is a sum of several factors to which the number of citations is a mere – and even a bad – proxy. Second, the method applied does not use any field or publication type normalization. The study does not endeavour to compare different scientific fields, but answer the core question of analysing the impact of research collaboration to the production

of excellence to the extent it is measured by citations. We do not sample data or attempt to introduce measures that would "correct" the biased or skewed nature of the whole population – rather we discuss the eventuality that latent variables impacting the results do exist.

Results

The overall volume of Finnish authored papers has increased substantially in the late 1990s, when the government increased national R&D funding, but has saturated to an average 3% growth subsequently. The average Finnish publication count growth in the 2000s is similar to that of the average growth of world scientific publications of 2% (Veugelers, 2010). Interestingly, this has had little or no impact on the percentage of cited papers, which has remained fairly constant throughout the time frame of this study (Table 1.). The concentration of citations to on average 62 % of publications is also in line with earlier findings and the overall decline in the concentration of citations (Larivière, Gingras, & Archambault, 2009).

Table 1. Summary of data used in the study.

Year	All papers	Cited papers % of all papers	Hot papers % from all papers	Other cited papers % from all papers	Non-cited papers % from all papers	Hot papers % of cited papers	Citation threshold for hot papers (FI FRC)	Hot papers' share of all citations received (FI FRC)
1995	4693	58,34 %	6,33 %	52,01 %	41,66 %	10,85 %	10	44,93 %
1996	5119	62,77 %	6,15 %	56,61 %	37,23 %	9,80 %	9,33	43,45 %
1997	6041	59,25 %	6,22 %	53,02 %	40,75 %	10,51 %	10	44,89 %
1998	8800	58,56 %	6,26 %	52,30 %	41,44 %	10,69 %	10	43,28 %
1999	8867	61,59 %	6,03 %	55,55 %	38,41 %	9,80 %	9,14	41,15 %
2000	9477	59,78 %	5,99 %	53,78 %	40,22 %	10,03 %	9,33	41,15 %
2001	9353	62,75 %	6,19 %	56,56 %	37,25 %	9,87 %	9,14	42,65 %
2002	9608	61,14 %	6,29 %	54,85 %	38,86 %	10,28 %	10	41,30 %
2003	9972	63,48 %	6,35 %	57,13 %	36,52 %	10,00 %	9,2	40,10 %
2004	10705	60,49 %	6,07 %	54,41 %	39,51 %	10,05 %	10	40,72 %
2005	10615	63,86 %	6,62 %	57,24 %	36,14 %	10,37 %	10	41,12 %
2006	11480	62,91 %	6,30 %	56,61 %	37,09 %	10,01 %	9,67	40,76 %
2007	11999	61,39 %	6,16 %	55,23 %	38,61 %	10,03 %	9,75	40,87 %
2008	12341	64,44 %	6,47 %	57,97 %	35,56 %	10.05%	10	41,85 %
2009	12484	65,14 %	6,71 %	58,43 %	34,86 %	10,31 %	9	42,49 %

NOTE: Hot papers selection based on FI FRC Citations.

The structure and nature of scientific excellence changes substantially when the citations received by papers are weighted with the share of institutional Finnish authors as opposed of using simply the whole count of citations. The average share of Finnish institutional authors in "*Hot papers"* based on "*FI FRC Citations"* is 87% between 1995 and 2009, whereas respective share for *"Hot*

papers" based on "*ABS Citations*" is 57%. During the period, the share of Finnish institutional authors declines about 10 percentage units in both datasets, but the difference persists throughout the period.

The fractional counting scheme used in this paper approaches authorship equally dividing the fractions to all authors as a simple division citations by the number of authors. A more elaborate method, based on for example a Hunt-type (Hunt, 1991) valuation of authorship, could yield different results. However, this is not in the scope of this study.

The features of *"Hot papers"* based on *"FI FRC Citations*" are relatively constant throughout the period. The citation threshold for the most cited 10% of all cited papers within the 4-year citation window doesn't move significantly, and the share that *"Hot papers"* capture from all citations remains relatively stable, being roughly 40%. (Table 1.) The most cited 1% of all cited papers capture around 10% of all "*FI FRC Citations"* received. The excellence is significantly more concentrated when "*Hot papers*" are selected by using "*ABS Citations"*, when the best 1% would capture around 15% of all citations received, but the "*Hot papers"* combined 45%.

The comparison of the basic features of *"FI Hot papers"* and *"ABS Hot papers"* alone confirms that if one doesn't control of the geographic location of institutional authors, assessments of research excellence based on citations are bound to include considerable amount of noise, blurring the focus of targeted policy instruments. This essentially because the Finnish research excellence would be considerably less "Finnish", and, secondly, excellence would concentrate more strongly in the most cited papers.

Structure and nature of research teams

The size of author teams is an essential factor when assessing the productivity (e.g. published papers or received citations). From the perspective of national research systems and policies, the nationality, and more accurately domestic and foreign authorship, are alike significant. Here we investigate how the average number of authors and the share of Finnish institutional authors evolve in the three data sets created with "*FI FRC Citations*". (Figure 1.)

The average number of authors in *"Hot papers"* and *"Non-cited papers"* is relative stable in 1995-2009 and not very different, being on average about 6 and 4, respectively. In contrast, *"Other cited papers"* shows a more challenging trend, with significant variance in the average number of authors, suggesting structural system-level transitions in the structure or focus of research, such as annual variation in participation to international big science. This is supported by the significantly smaller author count standard deviation of 9,4 in 2002 in comparison to the average of 45,7. Outliers in the data offer a reasonable explanation to the

behaviour, as when calculated as a median value the "Other cited papers" remains at a stable median of 4 authors until 2004. From 2005 onward, the median value of the group increases to 5 authors. This suggests that the upward trend from 2004 in the "Other cited papers" is an actual increase in group size. Disciplinary orientation could explain these differences too, but is beyond this paper. (Figure 1.)

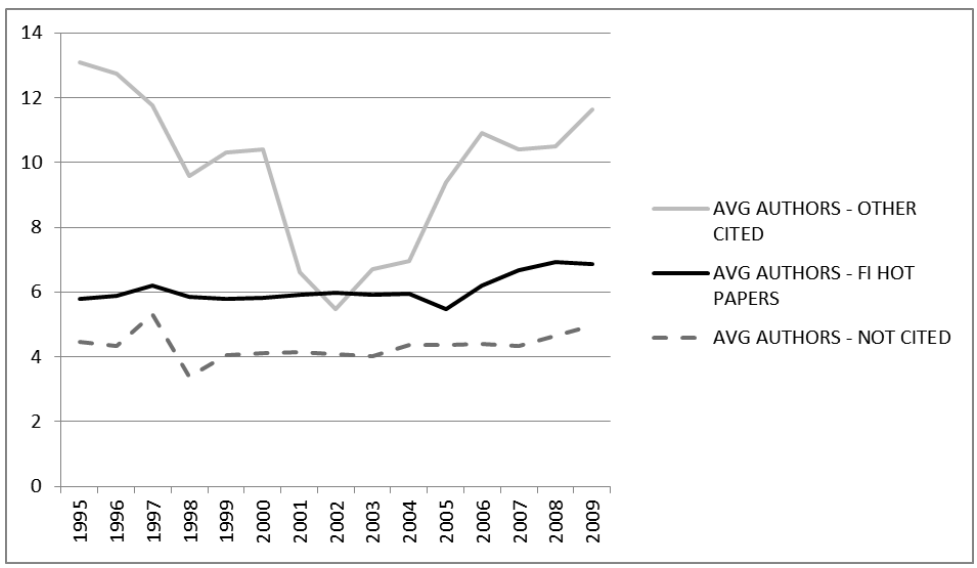


Figure 1. Average number of authors in the publications in the three groups (FI FRC Hot papers, FI FRC Other cited papers and Non-cited papers) during 1995 to 2009.

The share of Finnish institutional authors in *"Hot papers"* and *"Non-cited papers"* is almost equal throughout the period, being around 90% and, thus, highly "Finnish". In contrast, author teams for "*Other cited papers*" more and increasingly international, with the share of Finnish institutional authors declining steadily from about 80% in 1995 to around 65% in 2009. (Figure 2.)

When investigated from national research system perspective, the structure and nature of research teams in the three different quality groups of Finnish research is somewhat counter intuitive if one subscribes to the idea that large international teams author the most cited papers. The assumed best papers, *"Hot papers"*, are on average relatively small (6 authors) and have a very low degree of international collaboration (10% of institutional authors are non-Finnish). As such, they resemble closely the *"Non-cited papers"* that have about 4 authors on average and almost identically low degree of international collaboration. The author structure and national composition of these two groups remains also relatively stable between 1995 and 2009.

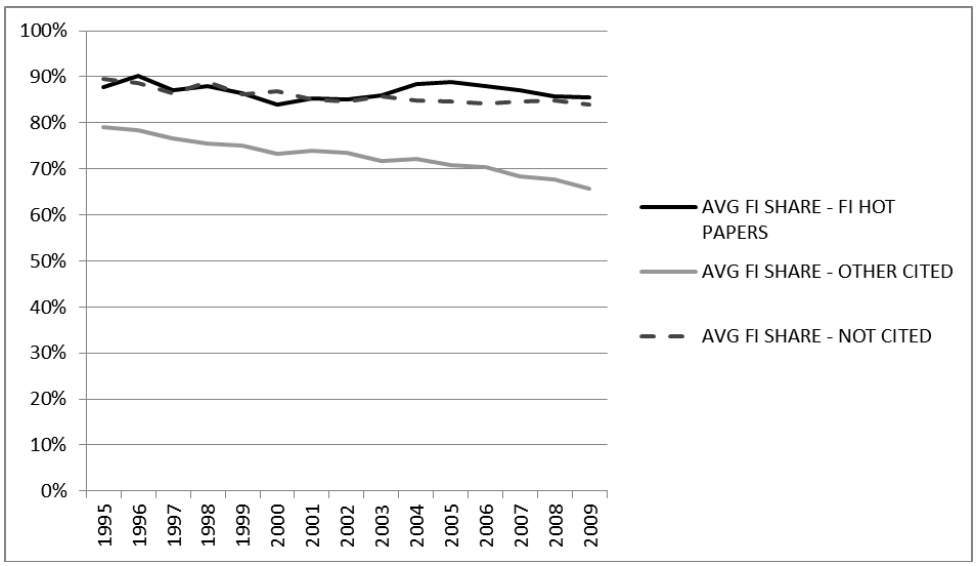


Figure 2. Average share of Finnish authors in the publications in the three groups (FI FRC Hot papers, FI FRC Other cited papers and Non-cited papers) during 1995 to 2009.

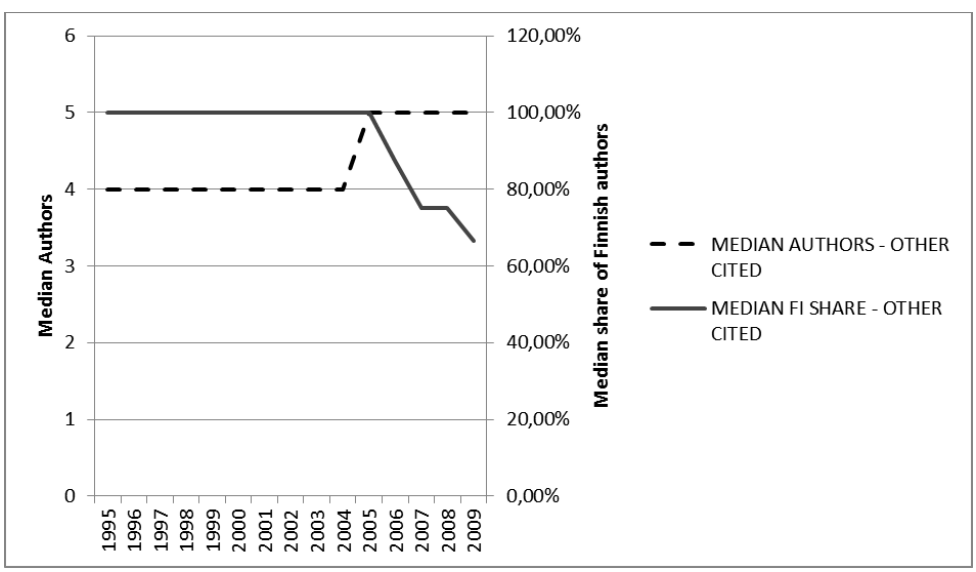


Figure 3. Median authors and share of Finnish authors in the "Other cited papers".

Remarkably, the *"Other cited papers"* shows very different qualities and trends. Its average author share is, for the most years, significantly higher than in the two other groups, and varies greatly. Furthermore, its author teams are significantly more and increasingly international, with the share of non-Finnish institutional

authors increasing from about 20% to 35%. This is further illustrated by the more stable median value, in Figure 3, where we see that as the author count increases the share of non-Finnish authors increase. While the selection criteria applied here, *"FI FRC Citations"*, explains these differences to some degree, as could the disciplinary orientation not afforded to study here.

Citation based excellence and productivity

To assess the overall citation-based quality and author productivity in the context of excellence, we compare the citations received by papers and author teams' productivity in attracting citations across the different data sets.

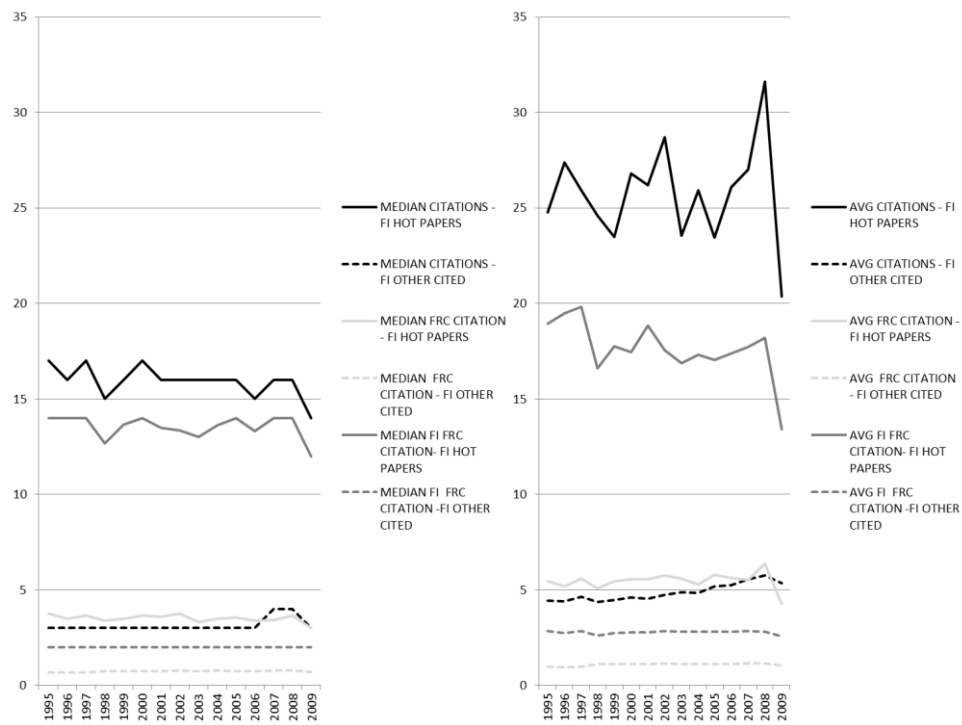


Figure 4. Average and median number of absolute citations (AVG/MEDIAN CITATIONS), average and median share of citations per author (AVG/MEDINA FRC CITATION) and average and median share of Finnish institutional authors from citations (AVG/MEDIAN FI FRC CITATION) within the two citation groups ("*Hot papers*" and "*Other cited papers*").

The different citation averages of "*Hot papers*", and "*Other cited papers*" (selected with the "*FI FRC Citation*") remain relatively stable throughout the period, yet the two groups have sharply contrasting features. As illustrated in Figure 4, the average (whole) citation value of "*Hot papers*" between 1995-2009 inches somewhat upwards, being on average 25 for the period, of which on

average about 17 can be credited to Finnish institutional authors - though this latter share declines somewhat during the period. (Note that 2009 citation count drops because the share of conference proceedings drops – something we will fix with data augmentation later).

In case of *"Other cited papers"* the average (whole) citation inches slightly up, being on average little below 5 between 1995 and 2009, of which Finnish institutional authors capture on average 2.7 citations. However, the "*FI Auth share*" declines from almost 80% to 65%.

Thus, the ratio of average citations and the share of Finnish institutional authors remain higher for the "*Hot papers*" than *"Other cited papers"* within the time period. It is noteworthy, that although the average number of authors in "*Other cited papers*" experiences a U-shaped curve, the average citation values remain as a constant throughout the period. This might be explained by the significantly smaller variance and the median being stable – the average paper performs similarly and the fluctuation in the number of authors does not change this. This however suggests that the so called "big science" papers, with an extremely large number of authors, perform near to average.

Figure 4. also includes the median value as a measure of central tendency. Although Aksnes and Sivertsen (2004) question the usefulness of median values in describing citations, even though the distribution of citations is prone to outliers and skewness, median values as a more stable measure of central tendency give useful insight to the average based measures. Looking at the median values, we see the same order of variables. In addition, the comparison between median and average values show that each of the variables have a similarly skewed distribution containing data points that are significantly larger than the central tendency of the distribution – e.g. all of the cohort are independently skewed.

To contrast the differences in results obtained by using national fractional citation counting and whole citation counting, we compare *"FI Hot papers"* and "*ABS Hot papers*" datasets to examine how they differ in terms of production of excellence. (Figure 5.)

The Figure 5. illustrates the fundamental problem addressed in our paper. Whereas "*ABS Hot papers"* receive by whole count more citations than *"FI Hot papers"*, and is by that definition more *"hot"*, the relationship is turned upside down when we use national fractional citation count. Indeed, with the "*FI FRC* citation" count, *"FI Hot papers"* is *"hotter*" than *"ABS Hot papers"*. This notion holds true in both average or median based metrics. With median values we even see that the gap between the "FI Hot papers" and *"ABS Hot papers"* has increased from a near equal.

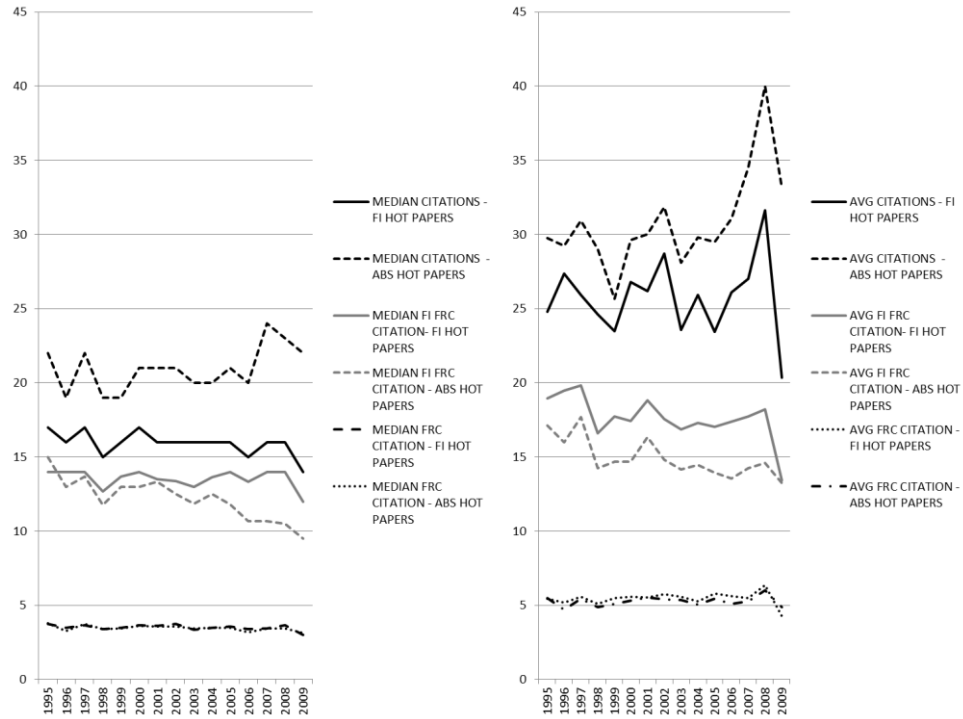


Figure 5. Comparison of citation profile between the "*FI Hot papers*" and "*ABS Hot papers*". Average whole count of citations (AVG/MEDIAN CITATIONS), average citations per author (AVG/MEDIAN FRC CITATION) and average share of citations of Finnish institutional authors from citations (AVG/MEDIAN FI FRC CITATION)

Perhaps the most important observation relates to the author productivity in terms of attracting citations. The fractional count of citations per author in *"FI Hot papers"* and *"ABS Hot papers"* is practically the same throughout 1995-2009, being roughly 5 (average 5.5 or median 3.5 and 5.3 or median 3.5, respectively). This despite the fact that the two groups differ greatly in terms of author team structure and the extent of international collaboration (and probably disciplinary orientation).

This last result raises the question what kind of excellence is important in the context of national research systems, and how excellence can be fostered with targeted policy instruments. Basically, our result contrasts here the assessment of excellence of papers with that of researchers, and our results show that highly national and relative small author teams are just as productive (and actually slightly more) in terms of scientific excellence as large international teams.

Discussion

By exploring the nature of Finnish research 1995-2009, we have demonstrated that the nature of the "*FI Hot papers"* or *research frontiers* doesn't correspond with the idealized vision of "high quality research". It is created by relatively small author teams (on average 6 co-authors) and is highly national (on average 85% of authors are Finnish), and that as such it resembles closely research with no impact, i.e. the *"Non-cited papers"* (4 authors; 86% Finnish, respectively). These two groups differ dramatically from the "*Other cited papers"* (10 authors, 73% Finnish, respectively).

More importantly, we have demonstrated that the *"Hot papers"* selected with *"FI FRC citation"* count are equally productive in terms of citations received per author as *"Hot papers"* selected with absolute citation count. This despite the fact that these two groups differ starkly, as the latter group has much larger author groups, and involve international collaboration to the extent that implies a marginal author role for Finnish contributors. In contrast, "*FI Hot papers"* are created by relatively small teams that are about 90% of Finnish, suggesting a leading and controlling role of Finnish contributors.

Methods developed here and our results raise fundamental questions about how to define scientific excellence in national research system context, and how to device policy strategies and targeted instruments in support of excellent researchers. Basically, which science is more valuable and potential from national perspective: The one created through participation in marginal role in international large collaborative author teams, or the one authored through participation in central leadership role in relatively small and highly national author teams?

While it is beyond this paper to address fully these questions, our results alone give rise to the suspicion that conventional (Finnish) policy efforts to foster research excellence target the middle-tier papers, and target poorly the best papers that resemble closely the worst ones, but is compounded when we show that the *"ABS Hot papers"* – the assumed crème of Finnish science - has less than half of Finnish institutional authorship.

As such, our results have immediate bearing upon public policies and institutional strategies trying to foster excellence in science, up to the point of suggesting that many of the existing approaches may be wrong-headed and target wrong (mediocre) researcher populations. This especially, if they are underpinned by simplistic assumptions that large and international author teams lead to scientific excellence, or are based on assessments of research that are blind to geographic sources of authorship. While our results may be a Finnish idiosyncry, the bibliometric assessment method and some literature (Toivanen, 2012b) suggests otherwise, as the fractional accounting of geographic locations shows great

variance in the share of domestic authorship. Naturally more comparative work is needed, as well as to develop further policy implications.

References

Aksnes, D., Schneider, J. W., & Gunnarsson, M. (2012). Ranking national research systems by citation indicators. A comparative analysis using whole and fractionalised counting methods. *Journal of Informetrics*, *6*(1), 36–43. doi:10.1016/j.joi.2011.08.002

Aksnes, D., & Sivertsen, G. (2004). The effect of highly cited papers on national citation indicators. *Scientometrics*, *59*(2), 213–224. Retrieved from http://www.akademiai.com/index/N167344L22728752.pdf

Beaver, D. D. B. (2001). Reflections on scientific collaboration (and its study): past, present, and future. *Scientometrics*, *52*(3), 365–377. Retrieved from http://www.akademiai.com/index/FA5LNKLHDKGD7X77.pdf

Bourke, P., & Butler, L. (1996). Standards issues in a national bibliometric database: The Australian case. *Scientometrics*, *35*(2), 199–207. doi:10.1007/BF02018478

Bozeman, B., Fay, D., & Slade, C. P. (2012). *Research collaboration in universities and academic entrepreneurship: the-state-of-the-art*. *The Journal of Technology Transfer* (Vol. 38, pp. 1–67). doi:10.1007/s10961-012-9281-8

Chao, C.-C., Yang, J.-M., & Jen, W.-Y. (2007). Determining technology trends and forecasts of RFID by a historical review and bibliometric analysis from 1991 to 2005. *Technovation*, *27*(5), 268–279. doi:10.1016/j.technovation.2006.09.003

Defazio, D., Lockett, A., & Wright, M. (2009). Funding incentives, collaborative dynamics and scientific productivity: Evidence from the EU framework program. *Research Policy*, *38*(2), 293–305. Retrieved from http://www.sciencedirect.com/science/article/pii/S0048733308002709

Edler, J., & Flanagan, K. (2011). Indicator needs for the internationalisation of science policies. *Research Evaluation*, *20*(1), 7–17.

Fraunhofer ISI, Idea Consult, & SPRU. (2009). *The Impact of Collaboration on Europe´s Scientific and Technological Performance, Final Report*. Karlsruhe, Brussels, Brighton.

Garfield, E. (2006). The history and meaning of the journal impact factor. *JAMA: the journal of the American Medical Association*, *295*(1), 90–93. Retrieved from http://jama.ama-assn.org/content/295/1/90.short

Garfield, E., & Merton, R. K. (1979). *Citation indexing: Its theory and application in science, technology, and humanities* (Vol. 8). Wiley New York. Retrieved from http://www.garfield.library.upenn.edu/cifwd.html

Garfield, E., & Small, H. (1989). Identifying the changing frontiers of science. *evolution*, *1*(182), 1431. Retrieved from http://www.garfield.library.upenn.edu/papers/362/362.html

Gauffriau, M., Larsen, P. O., Maye, I., Roulin-Perriard, A., & Von Ins, M. (2007). Publication, cooperation and productivity measures in scientific research. *Scientometrics*, *73*(2), 175–214.
Gauffriau, M., Larsen, P. O., Maye, I., Roulin-Perriard, A., & Von Ins, M. (2008). Comparisons of results of publication counting using different methods. *Scientometrics*, *77*(1), 147–176.
Glänzel, W., & Schoepflin, U. (1994). Discussion Paper LITTLE SCIENTOMETRICS , BIG SCIENTOMETRICS ... AND BEYOND ?*. *Scientometrics*, *30*, 375–384.
Glänzel, W., & Schubert, A. (2001). Double effort= double impact? A critical view at international co-authorship in chemistry. *Scientometrics*, *50*(2), 199–214. Retrieved from http://www.akademiai.com/index/T2PQ1372401V27V8.pdf
Hicks, D. (2012). Performance-based university research funding systems. *Research Policy*, *41*(2), 251–261. doi:10.1016/j.respol.2011.09.007
Ho, Y.-S. (2009). Comments on "Determining technology trends and forecasts of RFID by a historical review and bibliometric analysis from 1991 to 2005". *Technovation*, *29*(10), 725–727. doi:10.1016/j.technovation.2009.01.002
Ho, Y.-S. (2012). Comments on "a bibliometric study of earthquake research: 1900–2010". *Scientometrics*, (500), 2010–2012. doi:10.1007/s11192-012-0915-2
Huang, M. H., Lin, C. S., & Chen, D. Z. (2011). Counting methods, country rank changes, and counting inflation in the assessment of national research productivity and impact. *Journal of the American Society for Information Science and Technology*, *62*(12), 2427–2436.
Hunt, R. (1991). Trying an authorship index. *Nature*, *352*(6332), 187–187. Retrieved from http://old.imber.info/DM_WG/DM_portal/authorship.pdf
Ioannidis, J. P. A. (2008). Measuring co-authorship and networking-adjusted scientific impact. *PLoS One*, *3*(7), e2778. Retrieved from http://dx.plos.org/10.1371/journal.pone.0002778
Jones, B. F., Wuchty, S., & Uzzi, B. (2008). Multi-university research teams: shifting impact, geography, and stratification in science. *science*, *322*(5905), 1259–1262. Retrieved from http://www.sciencemag.org/content/322/5905/1259.short
Karlsson, S., & Persson, O. (2012). *The Swedish Production of Highly Cited Papers*. Vetenskapsrådet, Stockholm
Katz, J. S., & Hicks, D. (1997). How much is a collaboration worth? A calibrated bibliometric model. *Scientometrics*, *40*(3), 541–554. Retrieved from http://www.springerlink.com/index/M06J2871281PR764.pdf
Lange, L. (2001). Citation Counts of Multi-Authored Papers—First-named Authors and Further Authors. *Scientometrics*, *52*(3), 457–470. Retrieved from http://www.akademiai.com/index/Y5DM8J5A4Q90VMGA.pdf
Larivière, V., Gingras, Y., & Archambault, É. (2009). The decline in the concentration of citations, 1900–2007. *Journal of the American Society for*

Information Science and Technology, *60*(4), 858–862. Retrieved from http://onlinelibrary.wiley.com/doi/10.1002/asi.21011/full

Larsen, P. (2008). Commentary on: "Indicators of European public research in hydrogen and fuel cells—an input–output analysis". *International Journal of Hydrogen Energy*, *33*(4), 1455–1456. doi:10.1016/j.ijhydene.2007.12.003

Larsen, P. O. (2008). The state of the art in publication counting. *Scientometrics*, *77*(2), 235–251. doi:10.1007/s11192-007-1991-6

Lee, S., & Bozeman, B. (2005). The Impact of Research Collaboration on Scientific Productivity. *Social Studies of Science*, *35*(5), 673–702. doi:10.2307/25046667

Leydesdorff, L., & Opthof, T. (2010). Normalization at the field level: Fractional counting of citations. *arXiv preprint arXiv:1006.2896*. Retrieved from http://arxiv.org/abs/1006.2896

Liao, C. H. (2011). How to improve research quality? Examining the impacts of collaboration intensity and member diversity in collaboration networks. *Scientometrics*, *86*(3), 747–761. Retrieved from http://www.akademiai.com/index/X1724TP767401812.pdf

Liu, X., Zhan, F. B., Hong, S., Niu, B., & Liu, Y. (2012). A bibliometric study of earthquake research: 1900–2010. *Scientometrics*, *92*(3), 747–765. doi:10.1007/s11192-011-0599-z

McGrath, W. (1996). The unit of analysis (objects of study) in bibliometrics and scientometrics. *Scientometrics*, *35*(2), 257–264. Retrieved from http://www.springerlink.com/index/KN81081G526100LK.pdf

Moed, H. F. (2005). *Citation analysis in research evaluation* (Vol. 9). Springer. Retrieved from http://www.google.com/books?hl=en&lr=&id=D9SaJ6awy4gC&oi=fnd&pg=PR9&dq=Citation+analysis+in+research+evaluation&ots=FFnVIt3Rk_&sig=0m5FKrBb1wBHc9dnpddEoxu_NYI

Moed, H. F., De Bruin, R. E., & Van Leeuwen, T. N. (1995). New bibliometric tools for the assessment of national research performance: Database description, overview of indicators and first applications. *Scientometrics*, *33*(3), 381–422. Retrieved from http://www.akademiai.com/index/N0683864614TN345.pdf

Muhonen, R., Leino, Y., & Puuska, H.-M. (2012). *Suomen kansainvälinen yhteisjulkaiseminen*. Helsinki: Opetus- ja kulttuuriministeriö.

Nicholson, J. M., & Ioannidis, J. P. A. (2012). Research grants: Conform and be funded. *Nature*, *492*(7427), 34–36. doi:10.1038/492034a

Opetus- ja kulttuuriministeriö. (2011a). *Sitaatioindeksityöryhmä II : n raportti*. Opetus- ja kulttuuriministeriön julkaisuja 2011:34 (T*he Ministry of Education and Culture publications in Finnish), Helsinki*

Opetus- ja kulttuuriministeriö. (2011b). *Sitaatioindeksityöryhmän raportti*. Opetus- ja kulttuuriministeriön julkaisuja 2011:12 (T*he Ministry of Education and Culture publications in Finnish), Helsinki*

Persson, O., Glänzel, W., & Danell, R. (2004). Inflationary bibliometric values: The role of scientific collaboration and the need for relative indicators in evaluative studies. *Scientometrics*, *60*(3), 421–432. Retrieved from http://www.akademiai.com/index/h880j22v8t145572.pdf

Rons, N., & Amez, L. (2009). Impact vitality: an indicator based on citing publications in search of excellent scientists. *Research Evaluation*, *18*(3), 233–241. doi:10.3152/095820209X470563

Schmoch, U., & Schubert, T. (2008a). Are international co-publications an indicator for quality of scientific research? *Scientometrics*, *74*(3), 361–377. Retrieved from http://www.springerlink.com/index/8475PP504667561U.pdf

Schmoch, U., & Schubert, T. (2008b). Are international co-publications an indicator for quality of scientific research? *Scientometrics*, *74*(3), 361–377.

Seymour, E. H. (2008). Rebuttal to commentary. *International Journal of Hydrogen Energy*, *33*(4), 1457–1458. doi:10.1016/j.ijhydene.2007.12.004

Seymour, E. H., Borges, F. C., & Fernandes, R. (2007). Indicators of European public research in hydrogen and fuel cells—An input–output analysis. *International Journal of Hydrogen Energy*, *32*(15), 3212–3222. doi:10.1016/j.ijhydene.2007.02.031

Toivanen, H. (2012a). Identifying hot Brazilian science and technology: Tech mining methods for relating sources of knowledge and emerging research areas. Karlsruhe, Germany.

Toivanen, H. (2012b). Does the knowledge trap exist? The diverging roles of domestic and foreign capacities for the evolution of knowledge creation in China, India, Brazil, and Northern and Sub-Saharan Africa. Hangzhou, China.

Treudthardt, L., & Nuutinen, A. (Eds.). (2012). *Suomen tieteen tila*. Helsinki: Suomen Akatemia.

Wagner, C. S., & Leydesdorff, L. (2005). Mapping the network of global science: comparing international co-authorships from 1990 to 2000. *International Journal of Technology and Globalisation*, *1*(2), 185–208. Retrieved from http://inderscience.metapress.com/index/q8x345099crntewm.pdf

Wagner, C. S., & Leydesdorff, L. (2012). An Integrated Impact Indicator: A new definition of "Impact"with policy relevance. *Research Evaluation*, *21*(3), 183–188. Retrieved from http://rev.oxfordjournals.org/content/21/3/183.short

Waltman, L., Van Eck, N. J., Van Leeuwen, T. N., Visser, M. S., & Van Raan, A. F. J. (2011). Towards a new crown indicator: Some theoretical considerations. *Journal of Informetrics*, *5*(1), 37–47. Retrieved from http://www.sciencedirect.com/science/article/pii/S1751157710000817

Veugelers, R. (2010). Towards a multipolar science world: trends and impact. *Scientometrics*, *82*(2), 439–456. Retrieved from http://www.akademiai.com/index/b60146401036v185.pdf

Vinkler, P. (1996). Some practical aspects of the standardization of scientometric indicators. *Scientometrics*, *35*(2), 237–245. doi:10.1007/BF02018481

Vinkler, P. (2001). An attempt for defining some basic categories of scientometrics and classifying the indicators of evaluative scientometrics.

Scientometrics, *50*(3), 539–544. Retrieved from http://www.springerlink.com/index/q31h04230660123p.pdf

Zhao, D. (2006). Dispelling the Myths Behind First-author Citation Counts. *Proceedings of the American Society for Information ...*, *43*(1), 1–16. Retrieved from http://onlinelibrary.wiley.com/doi/10.1002/meet.14504301194/full

RESEARCH PERFORMANCE ASSESSMENT USING NORMALIZATION METHOD BASED ON SCI DATABASE (RIP)

Ling Zhang[1], Juan Wang[2], Yanan Zhang[3], Xin Tan[4], Qing Du[5]

[1]*lingzhang@tju.edu.cn*
Tianjin University, 92 Weijin Road, Nankai District, Tianjin (China)

[2]*rscwangjuan@tju.edu.cn*
Tianjin University, 92 Weijin Road, Nankai District, Tianjin (China)

[3]*zhangnanmiao@gmail.com*
Tianjin Polytechnic University, extension line, BinshuiXi Road, Xiqing District, Tianjin (China)

[4]*tanx@tju.edu.cn*
Tianjin University, 92 Weijin Road, Nankai District, Tianjin (China)

[5]*duqing@tju.edu.cn*
Tianjin University, 92 Weijin Road, Nankai District, Tianjin (China)

Abstract
To correct differences among fields, a derivative indicator of Crown Indicator - T-indicator - was proposed as an effective supplement of established Article Assessment System of Tianjin University. Based on normalized citation counts, T-indicator could give the order of research performance of researchers or groups in different disciplines. A given example was used to thoroughly discuss this evaluation method, via the application of derivative indices using SCI database.

Conference Topic
Scientometrics Indicators (Topic 1) and Science Policy and Research Evaluation: Quantitative and Qualitative Approaches (Topic 3)

Introduction
Research performance assessment (RPA) plays important roles in universities and research institutions, especially in the process of recruitment, academic promotion, offering tenure, granting, etc. The general indices of RPA include publications, patents, awards, and grants. It is hard to evaluate the quality level of patents, awards, and grants among different institutions and countries as there is no same standard. However, journal publication, mostly published after peer reviews, is a good and unique index for internal and external comparison.

Nowadays, journal publication has been widely used officially or subconsciously in the process of RPA.

An article assessment system has been successfully established based on both Tianjin University and nine key Chinese Universities' academic disciplinary benchmarks (Zhang[a], 2010). With this scientific benchmarking system, the quality of a researcher's papers could be easily located in a percentile scale in corresponding field and within certain groups. Several factors, including total number of papers, order of authors, impact factor of journals, citation count, h-index (Hirsch, 2005), e-index (Zhang[b], 2009), a-index (Jin, 2006), m-quotient (Hirsch, 2005), as well as weighted citation analysis (Zhang[c], 2009), were also utilized for both quantity and quality analysis.

This article assessment system has played a significant role as an important part of RPA in Tianjin University. However, with unique advantages in comparing researchers or groups in a same field, it is hard to tell their RPA in different fields. To improve this article assessment system, referring to Crown Indicator proposed by CWTS, a derivative indicator, named after Tianjin University - T-indicator - was applied, where citation counts were normalized for correcting differences among fields. (Zhang[d], 2012) However, in this early study of T-indicator, the applied disciplines were only 25 categories defined in the Scopus database, which are very broad for building subject-specific indicators and reference standards. In this paper, 169 disciplines based on SCI database were applied. The modified citation-based article assessment system could easily and specifically give the order of research performance of researchers or groups even in different disciplines.

Method

The average number of citation count of all TJU publications from SCI citation database are obtained for each discipline and for each year from the year of 2001 to the year of 2011, based on the accumulation of citations from the year of publication to the current year. (Equation 1)

$$AC_{y_j} = \frac{1}{n_j} \sum_{i=1}^{n_j} C_{i,j} \qquad (1)$$

where $C_{i,j}$ are the citations received by the i^{th} paper in the year j, and n_j is number of papers published in the year j. On the left hand of equation (1), AC_{yj} represents the average number of citations received in the period from year j to 2011 by papers published in the year j.

To obtain the total T–indicator (T_{total}), annual T–indicator (T_{year}) are required to be calculated firstly: the sum of a researcher or group's actual number of citations of all publications is divided by the above average number for each year in the same discipline (Equation 2).

$$T_{y_j} = \frac{1}{m_j} \frac{\sum_{i=1}^{m_j} C_{i,j}}{AC_{y_j}} \qquad (2)$$

where m_j is the number of papers published by an individual researcher or a group of researchers in the year j, and T_{yi} is the ratio of the average citations received for an individual researcher or a group of researchers in the year j, over the average number of citations received in the year j of the whole university, both in the in the same discipline.

The average number of T_{year} is the T-indicator (Equation 3)

$$T_{total} = \frac{1}{(y_2 - y_1 + 1)} \sum_{j=y_1}^{y_2} TJ_{y_j} \qquad (3)$$

Where y_1 is the first year of the period in which the research performance of an individual researcher or a group of researchers are required to be analyzed, and y_2 is last year of this period required to be analyzed.

Results and Discussion

Table of *Mean of Citation Count of all TJU Publications* is prepared (Table 1) for 169 disciplines from the year of 2001 to the year of 2011. Total number of TJU publications over 11 years and of each year, as well as the annual mean citation count were all included for every category in this table. For example, in category of "ENGINEERING CHEMICAL", total number of TJU publication is 1587; the number of publications in the year of 2001 and the mean citation count is 41 and 9.8, respectively.

Table 1. The Mean of Citation Count of all TJU Publications. The data were collected from SCI citation database at Oct.2012. (The table is too big to present entirely.)

No.	*Subject*	*Total*	*2001*		*2002*		*2003*	
		Pub.	*Pub.*	*Mean*	*Pub.*	*Mean*	*Pub.*	*Mean*
1	MATERIALS SCIENCE MULTIDISCIPLINARY	1731	36	4.0	57	7.8	71	8.4
2	ENGINEERING CHEMICAL	1587	41	9.8	91	9.5	96	11.0
3	CHEMISTRY PHYSICAL	1356	44	15.9	58	9.5	72	9.0
4	CHEMISTRY MULTIDISCIPLINARY	1164	31	5.3	37	8.1	64	6.9
5	PHYSICS APPLIED	1009	12	14.1	21	8.9	32	20.2
…	…	…	…	…	…	…	…	…
169	UROLOGY NEPHROLOGY	1	0	0	0	0	0	0

The following example is taken to discuss the application of T-indicator. Tianjin University announced the competition for an award funding for research performance, and there are 8 candidates. In the process of research publication assessment, as shown in Table 2, all of them are excellent in their research fields, and some of them have similar number of publications (Candidate 5 and Candidate 6), total citation count (Candidate 1 and Candidate 6), and average citation count (Candidate 2 and Candidate 3) as well. Furthermore, considering the property of citation frequency in different research areas, it is very hard to simply compare them via the common indices, including citation count, h-index, e-index, etc., as mentioned above. However, T-indicator, based on normalized citation count, could be conveniently used here to give the order of research performance as a helpful reference to the award funding committee.

Table 2. Publication details of 8 candidates for the award funding for research performance. The data were collected from SCI citation database at Oct.2012.

No.	*College*	*Total pub.*	*Total citation count*	*Average citation count*
1	College of Science	130	1079	8.3
2	College of Science	344	5564	16.2
3	College of Science	54	924	17.1
4	College of Precision Instrument and Opto-electronics Engineering	57	368	6.5
5	College of Precision Instrument and Opto-electronics Engineering	101	891	8.8
6	College of Material Science and Engineering	111	1051	9.5
7	College of Chemical Engineering and Technology	69	979	14.2
8	College of Environment Science and Technology	159	2999	18.9

In SCI citation database, collected journals are categorized into 169 disciplines; however, due to the relativity among certain fields, publications of some journals are subjected to 2 or even more disciplines. In such case, the average of T-indicators of different disciplines could be used instead, due to the normalized native of T-indicator. For example, Candidate 1 has published 130 articles, which are categorized to 12 disciplines by SCI, including “Physics Applied” (65), “Physics Condensed Matter” (48), “Materials Science Multidisciplinary” (34), and so on. Apparently some of the publications are classified to several disciplines by SCI, instead of only one category.

As shown in Table 3, in Discipline 1 - the category of “Physics Applied”, averages of citation count of different year were calculated firstly (Row 3), which were then divided by the corresponding average number of citation count of all TJU publications for each year in Table 1, and the quotients—obtained (Row 4) were T_{year}-indicator. T_1 (0.77) of “Discipline 1” was subsequently calculated. The

same method was also used to calculate the T_2 (0.51) of publications in Discipline 2 of "Physics Condensed Matter". With this method, T indicator could be obtained respectively for the next 10 disciplines. Finally T_{total} (0.42) was achieved by computing the mean value of them in different subjects.

Table 3. T-indicator of publication of Candidate 1. The data were collected from SCI citation database at Oct. 2012.

Candidate 1											
Discipline 1: Physics Applied											
Year	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011
No. of Pub.	2	3	6	5	5	10	8	9	7	7	3
No. of citation count	49	7	104	68	44	55	42	61	23	18	2
Aver. of citation count	24.5	2.3	17.3	13.6	8.8	5.5	5.2	6.8	3.3	2.6	0.7
T_{year}	1.74	0.26	0.86	1.27	0.83	0.54	0.63	0.60	0.55	0.69	0.49
T_1=0.77											
Discipline 2: Physics Condensed Matter											
Year	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011
No. of Pub.	3	4	5	7	3	7	7	3	6	2	1
No. of citation count	4	131	15	49	14	42	32	3	20	1	0
Aver. of citation count	1.3	32.8	3.0	7.0	4.7	6.0	4.6	1.0	3.3	0.5	0.0
T_{year}	0.35	1.95	0.27	0.86	0.39	0.57	0.48	0.07	0.47	0.17	0.00
T_2=0.51											
Discipline … …											
… …											
Discipline 12: Metallurgy Metallurgical Engineering											
Year	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011
No. of Pub.	1	1	2	1	2	1	3	2	0	1	0
No. of citation count	0	0	13	16	33	0	25	5	0	2	0
Aver. of citation count	0.0	0.0	6.5	16.0	16.5	0.0	8.3	2.5	0.0	2.0	0.0
T_{year}	0.00	0.00	3.20	5.90	3.57	0.00	2.28	0.77	0.00	0.63	0.00
T_{12}=1.49											
T_{total}=0.42											

T_{total} could also show an individual annual research performance as shown in Figure 1. For example, T_{total} of Candidate 1 hit the peak (1.47) in the year of 2004, and reached the bottom after the year of 2008, presenting a decreasing research performance. However, T_{total} of Candidate 2 has gradually climbed up since the year of 2003, and a sudden jump to the peak of 1.22 appeared in the Year of 2004, after that a stable trend was shown, demonstrating an increasing research performance. A conclusion could be drawn that both Candidate 1 and 2 are very excellent in their own research field as their T_{total} are almost over 1, but Candidate 2 showed higher potential in research.

When comparing the research performance among more scholars in different disciplines, T_{total} displays unique advantages. As shown in Table 4, T_{total} of each

candidate was calculated, and from these data, Candidate 8 showed the best research performance with the highest T_{total} of 1.69, followed by Candidate 5 and Candidate 4, with 1.13 and 0.94, respectively, and the poorest performance in this group is Candidate 1, showing the lowest T_{total} of 0.42.

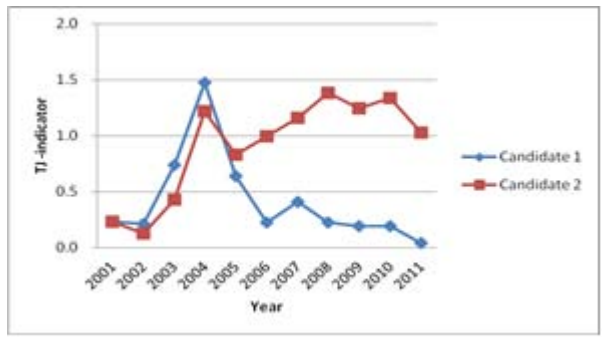


Figure 1. T_{total} vs. year of Candidate 1 and Candidate 2. (The data were collected from SCI citation database at Oct. 2012.)

For further analysis when considering candidates' contributions to publications, weighted T is introduced based on weighted citation analysis. The use of weighted citation analysis has been thoroughly discussed elsewhere (Zhang[c], 2009; Zhang[a], 2010), which is a quantitative scheme to describe the contribution of co-authors via weight coefficient. Basically weight coefficients for the first and corresponding authors are 1 for both, and the correspondence of the second, third, and the other authors are decreased sequentially. Weighted T of each candidate was obtained in Table 5. The weighted T-indicator was very similar to the normal T-indicator of both Candidates 3 (0.34 and 0.33, respectively), showing his/her high research contributions to all publications; however, the big difference of these two indicators of Candidate 4 (0.94 and 0.26, respectively) demonstrated his/her un-ideal contribution to all publications. Consequently, the order of research performance of these candidates based on weighted T-indicator could be listed as Candidate 8, Candidate 2, Candidate 7, Candidate 5, Candidate 3, Candidate 6, Candidate 4, and Candidate 1, without the consideration of differences among disciplines.

As described above, the research performance of these 8 candidates was quantitatively analyzed via this assessment method, which could give helpful reference to the award funding committee but still need the comprehensive qualitative evaluation *via* peer reviews, to get a final reasonable evaluation result of research performance of these candidates.

Table 4. T_{total} of publication of 8 candidates. The data were collected from SCI citation database at Oct. 2012.

No.	*Field No.*	T_{total}
1	12	0.42
2	26	0.91
3	22	0.34
4	15	0.94
5	8	1.13
6	17	0.48
7	13	0.93
8	19	1.69

Table 5. Weighted T_{total} of publication of 8 candidates. The data were collected from SCI citation database at Oct. 2012.

No.	*Field No.*	T_{total}
1	12	0.12
2	26	0.57
3	22	0.33
4	15	0.26
5	8	0.42
6	17	0.28
7	13	0.53
8	19	1.16

Conclusions

The application of T-indicators has successfully corrected differences among disciplines during research performance assessment using SCI database. An example was given to describe this whole assessment procedure which could not only give the research performance curve with year of candidate, but also provide the order of their research performance. Last but not least, because of the increasing citation times with time, the Table of the Mean of Citation Count of all TJU Publications is required to be updated at least twice annually.

Acknowledgments

We thank Prof. Chun-Ting Zhang and Prof. Qiushi Li of Tianjin University for helpful discussions and revisions. Ling Zhang thanks the financial support from The Ministry of education of Humanities and Social Science Research Fund Plan/Youth Fund/Self-financing project (11YJC870036) as well as Open Fund IT2012006 of the ISTIC-Thomson Reuters Joint Lab for Scientometrics Research.

References and Citations

Zhang[a], L., Zhao, H., Li, Q., Wang, J. & Tan, X. (2010) Establishment of Paper assessment system based on acadmic disciplinary benchmarks. *Scientometrics*, 84, 421-429.

Hirsch, J.E. (2005) An index to quantify an individual's scientific research output. *Proceedings of the National Academy of Sciences of the United States of America*, 102, 16569-16572.

Zhang[b], C.T. (2009) The e-index, complementing the h-index for excess citations. *PLoS (Public Library of Science) ONE*, 4, e5429.

Jin, B. (2006) h-index: an evaluation indicator proposed by scientist. *Science Focus*, 1, 8–9.

Zhang[c], C.T. (2009) A proposal for calculating weighted citations based on author rank. *EMBO (European Molecular Biology Organization) Reports*, 10, 416-417.

Zhang[d,] L., Tan, X., Du, Q., Wang, J. (2012) Research Performance Assessment based on T-indicator. Open Journal of Statistics, 2, 46-351.

RETHINKING RESEARCH EVALUATION INDICATORS AND METHODS FROM AN ECONOMIC PERSPECTIVE: THE FSS INDICATOR AS A PROXY OF PRODUCTIVITY

Giovanni Abramo[a, *], Ciriaco Andrea D'Angelo[a,b]

[a] Laboratory for Studies of Research and Technology Transfer
Institute for System Analysis and Computer Science (IASI-CNR)
National Research Council of Italy
Viale Manzoni, 30; 00185 Roma - Italy

[b] School of Engineering, Department of Management
University of Rome "Tor Vergata"

Abstract
The current bibliometric indicators and methods for evaluation of research performance are generally inappropriate in the light of economic theory of production. World ranking lists, including those by noted research agencies, seem based on what can be easily counted rather than "what really counts". In this work we operationalize the economic concept of productivity for the specific context of research activity and propose a measurable form of productivity. From an economic perspective, we demonstrate the limits of the most commonly-used performance indicators and we present the indicator "Fractional Scientific Strength (FSS)", which better approximates the measure of research productivity. We present the methodology for measure of FSS at various levels of analysis: individual, field, discipline, department, institution, region and nation.

Conference Topic
Topic 1 - Scientometrics Indicators: Criticism and new developments

Introduction
In 2010, Opthof & Leydesdorff criticized the statistical normalization of the Leiden CWTS "crown indicator". A year later, bibliometricians from the CWTS group (Waltman et al., 2011) admitted that the "old crown indicator" was mathematically inconsistent and adopted the normalization method suggested by the above authors, leading to a "new crown indicator": the mean normalized citation score, or MNCS. A counter-reply from Leydesdorff & Opthof (2011) was not long in arriving: although agreeing with the new statistical normalization, they then further recommended using the mean rather than the median to field normalize citations.

In a parallel story, since the original introduction of the h-index in 2005 by physicist Jorge E. Hirsch, over 1,300 articles have been written illustrating its merits and defects and proposing one variant after another, to the extent even the most devoted historian of bibliometrics would despair of tracing them all.

But is it possible that these two indicators really merited all this attention, or is it a case of "Much ado about nothing"? These particular indicators have only been the most popular among a myriad of others proposed over recent years by scholars and practitioners. While bibliometricians undoubtedly intended to provide useful indicators and ever more accurate and reliable methods, they have actually been the cause of increasing confusion. The proliferation of proposals has actually generated a type of disorientation among decision makers, no longer able to discriminate the pros and cons of the various indicators for planning an actual evaluation exercise. The proof of this is the increasing number of expert commissions and working groups at institutional, national and supranational levels, formed to deliberate and recommend on this indicator, that set of indicators, and this or that measure of performance. Performance ranking lists at national and international levels are published with media fanfare, influencing opinion and practical choices. The impression of the current authors is that these rankings of scientific performance, produced by "non-bibliometricians" (THES, 2012; ARWU, 2011; QS, 2011; etc.) and even by bibliometricians (University of Leiden, SCImago, etc.), are largely based on what can easily be counted rather than "what really counts". It is also our impression that the large part of the performance evaluation indicators proposed in the literature arise from a primarily mathematical school of thought. While knowledge in this area is fundamental in the methodology for application, our personal conviction is that research evaluation indicators must necessarily derive from economic theory. Since research activity is a production process, it should be analyzed from the perspective of microeconomic theory of production. Performance, or the ability to perform, should be evaluated with respect to the specific goals and objectives to be achieved. The objectives must therefore be stated in measurable terms representing the desired outcome of production activity. The principal performance indicator of a production unit (whether this an individual, research group, department, institution, field, discipline, region or country) is its productivity, or simply speaking the ratio of the value of the production output to the value of the inputs required to produce it. From this point of view, we will see that the renowned crown indicator and h-index, with its innumerable variants, and a wide variety of other publication-based and citation-based indicators, are inadequate to measure research productivity. As a consequence, all the research evaluations based on these indicators and their relative rankings are at best of little or no value, and are otherwise actually dangerous, due to the distortions embedded in the information provided to the decision-makers. For the large part of the objectives and contexts where evaluation of research performance is conducted, productivity is either the most important or the only indicator that should inform policy, strategy and operational decisions. We thus issue a two-fold

call to the scholars in the subject: first, to focus their knowledge and skills on further refining the measurement of this indicator in contexts of real use; second, to refrain from distribution of institutions' performance ranking lists based on invalid indicators, which could have negative consequences when used by policy-makers and research administrators.
In this work we intend to operationalize the concept of productivity for the specific context of research activity and propose a measurable form of productivity. We will then present an indicator, Fractional Scientific Strength (FSS), which in our view is thus far the best in approximating the measure of productivity. We will also illustrate the methodology for measuring FSS in the evaluation of performance at various levels of analysis: individual, field, discipline, department, institution, region and nation.

Productivity in research activities

In this section, our intention is to operationalize the concept of research productivity in simple terms and propose a proxy to measure it.
Generally speaking, the objective of research activity is to produce new knowledge. Research activity is a production process in which the inputs consist of human, tangible (scientific instruments, materials, etc.) and intangible (accumulated knowledge, social networks, economic rents, etc.) resources, and where output, the new knowledge, has a complex character of both tangible nature (publications, patents, conference presentations, databases, etc.) and intangible nature (tacit knowledge, consulting activity, etc.). The new-knowledge production function has therefore a multi-input and multi-output character. The principal efficiency indicator of any production unit (individual, research group, department, institution, field, country) is productivity, i.e. the ratio of the value of output produced in a given period to the value of production factors used to produce it. To calculate research productivity one needs adopt a few simplifications and assumptions.
On the output side, a first approximation arrives from the imposition of not being able to measure any new knowledge that is not codified. Second, where new knowledge is indeed codified, we are faced with the problem of identifying and measuring its various forms. It has been shown (Moed, 2005) that in the so-called hard sciences, the prevalent form of codification for research output is publication in scientific journals. Such databases as Scopus and Web of Science (WoS) have been extensively used and tested in bibliometric analyses, and are sufficiently transparent in terms of their content and coverage. As a proxy of total output in the hard sciences, we can thus simply consider publications indexed in either WoS or Scopus[142]. With this proxy, those publications that are not censused will inevitably be ignored. This approximation is considered acceptable in the hard

[142] Although the overall coverage of the two databases does differ significantly, evidence suggests that, with respect to comparisons at large scale level in the hard sciences, the use of either source yields similar results (Archambault et al., 2009).

sciences, although not for the arts, humanities and a good part of the social science fields. Other forms of output, particularly patents, can be identified in commercial or free databases such as Derwent and Espacenet. Patents are often followed by publications that describe their content in the scientific arena, so the analysis of publications alone may actually avoid in many cases a potential double counting.

Research projects frequently involve a team of researchers, which shows in co-authorship of publications. Productivity measures then need to account for the fractional contributions of single units to outputs. The contributions of the individual co-authors to the achievement of the publication are not necessarily equal, and in some fields the authors signal the different contributions through their order in the byline. The conventions on the ordering of authors for scientific papers differ across fields (Pontille, 2004; RIN, 2009), thus the fractional contribution of the individuals must be weighted accordingly. Following these lines of logic, all performance indicators based on full counting or "straight" counting (where only the first author or the corresponding author receive full credit and all others receive none) are invalid measures of productivity. The same invalidity applies to all indicators based on equal fractional counting in fields where co-author order has recognized meaning.

Furthermore, because the intensity of publications varies across fields (Garfield, 1979; Moed et al., 1985; Butler, 2007), in order to avoid distortions in productivity rankings (Abramo, D'Angelo & Di Costa, 2008), we must compare organizational units within the same field. A prerequisite of any productivity assessment free of distortions is then a classification of each individual researcher in one and only one field. An immediate corollary is that the productivity of units that are heterogeneous for fields of research of their staff cannot be directly measured at the aggregate level, and that there must be a two-step procedure: first measuring the productivity of the individual researchers in their field, and then appropriately aggregating this data.

In bibliometrics we have seen the evolution of language where the term "productivity" measures refers to those based on publication counts while "impact" measures are those based on citation counts. In a microeconomic perspective, the first operational definition would actually make sense only if we then compare units that produce output of the same value. In reality this does not occur, because the publications embedding the new knowledge produced have different values. Their value is measured by their impact on scientific advancements. As proxy of impact bibliometricians adopt the number of citations for the units' publications, in spite of the limits of this indicator (negative citations, network citations, etc.) (Glänzel, 2008). Citations do in fact demonstrate the dissemination of knowledge, creating conditions for knowledge spillover benefits. Citations thus represent a proxy measure of the value of output.

Comparing units' productivity by field is not enough to avoid distortions in rankings. In fact citation behavior too varies across fields, and is not unlikely that researchers belonging to a particular scientific field may also publish outside that

field (a typical example is statisticians, who may apply statistics to medicine, physics, social sciences, etc.). For this reason we need to standardize the citations of each publication with respect to a scaling factor stemming from the distribution of citations for all publications of the same year and the same subject category.[143] Different scaling factors have been suggested and adopted to field normalize citations (average, median, z-score of normalized distributions, etc.).
On the side of production factors, there are again difficulties in measure that lead to inevitable approximations. The identification of production factors other than labor and the calculation of their value and share by fields is not always easy (consider quantifying value of accumulated knowledge or scientific instruments shared among units). Furthermore, depending on the objectives of the assessment exercise, it could sometimes be useful to isolate and examine the contribution to output of factors, that are independent of the capacities of the staff for the units under examination (for example returns to scale, returns to scope, available capital, etc.).

Labor productivity in research activity and the FSS

The productivity of the total production factors is therefore not easily measurable. There are two traditional approaches used by scholars to measure the total factor productivity: parametric and non-parametric techniques. Parametric methodologies are based on the a priori definition of the function that can most effectively represent the relationship between input and output of a particular production unit. The purpose of non-parametric methods, on the other hand, is to compare empirically measured performances of production units (commonly known as Decision Making Units, DMUs), in order to define an "efficient" production frontier, comprising the most productive DMUs. The reconstruction of that frontier is useful to assess the inefficiency of the other DMUs, based on minimum distance from the frontier.
The measure of total factor productivity requires information on the different production factors by unit of analysis. Instead of total factor research productivity, most often research administrators are interested in measuring and comparing simply labor productivity, i.e. the value of output per unit value of labor, all other production factors being equal. In measuring labor productivity then, if there are differences of production factors available to each unit, one should normalize for these. Unfortunately, relevant data are not easily available, especially at the individual level. Thus an often-necessary assumption is that the resources available to units within the same field are the same. A further assumption, again unless specific data are available, is that the hours devoted to research are more or less the same for each individual. Finally, the cost of labor is likely to vary among research staff, both within and between units. In a study of Italian universities, Abramo, D'Angelo & Di Costa (2011) demonstrated that productivity of full,

[143] The subject category of a publication corresponds to that of the journal where it is published. For publications in multidisciplinary journals the scaling factor is generally calculated as the average of the standardized values for each subject category.

associate and assistant professors is different. Because academic rank determines differentiation in salaries, if information on individual salaries in unavailable then one can still reduce the distortion in productivity measures by differentiating performance rankings by academic rank.

Next we propose our best proxy for the measurement of the average yearly labor productivity at various unit levels (individual, field, discipline, department, entire organization, region and country). The indicator is named "Fractional Scientific Strength" (FSS), and we have previously applied it to the Italian higher education context, where most of its embedded approximations and assumption are legitimate.

As noted above, for any productivity ranking concerning units that are non-homogenous for their research fields, it is necessary to start from the measure of productivity of the individual researchers or fields. Without these two building blocks, any measure at aggregate level presents strong distortions (Abramo, D'Angelo & Di Costa, 2008). In their measures of this data, the authors gain advantage from a characteristic that seems unique to the Italian higher education system, in which each professor is classified as belonging to a single research field. These formally-defined fields are called "Scientific Disciplinary Sectors" (SDSs): there are 370 SDSs, grouped into 14 "University Disciplinary Areas" (UDAs). In the hard sciences, there are 205 such fields[144] grouped into in nine UDAs.[145]

When measuring research productivity, the specifications for the exercise must also include the publication period and the "citation window" to be observed. The choice of the publication period has to address often contrasting needs: ensuring the reliability of the results issuing from the evaluation, but also permitting conduct of frequent assessments. For the most appropriate publication period to be observed see Abramo, D'Angelo & Cicero (2012a), while for the citation window that optimizes the tradeoff between accuracy of rankings and timeliness of the evaluation exercise, see Abramo, D'Angelo & Cicero (2012b).

Labor productivity at the individual level

At micro-unit level (the individual researcher level, *R*) we measure Fractional Scientific Strength (FSS_R), a proxy of the average yearly productivity over a period of time, accounting for the cost of labor. In formula:

$$FSS_R = \frac{1}{s} \cdot \frac{1}{t} \sum_{i=1}^{N} \frac{c_i}{\bar{c}_i} \mathrm{f_i} \qquad [1]$$

Where:

s = average yearly salary of the researcher

t = number of years of work of the researcher in the period of observation;

N = number of publications of the researcher in the period of observation;

[144] The complete list is accessible on http://attiministeriali.miur.it/UserFiles/115.htm

[145] Mathematics and computer sciences; physics; chemistry; earth sciences; biology; medicine; agricultural and veterinary sciences; civil engineering; industrial and information engineering.

c_i = citations received by publication i;
$\bar{c}_i$ = average of the distribution of citations received for all cited publications[146] of the same year and subject category of publication i;
f_i = fractional contribution of the researcher to publication i.
Fractional contribution equals the inverse of the number of authors, in those fields where the practice is to place the authors in simple alphabetical order, but assumes different weights in other cases. For the life sciences, widespread practice in Italy and abroad is for the authors to indicate the various contributions to the published research by the order of the names in the byline. For these areas, we give different weights to each co-author according to their order in the byline and the character of the co-authorship (intra-mural or extra-mural). If first and last authors belong to the same university, 40% of citations are attributed to each of them; the remaining 20% are divided among all other authors. If the first two and last two authors belong to different universities, 30% of citations are attributed to first and last authors; 15% of citations are attributed to second and last author but one; the remaining 10% are divided among all others[147].
To calculate productivity accounting for the cost of labor, requires knowledge of the cost of each researcher, information that is usually unavailable for reasons of privacy. In the Italian case we have resorted to a proxy. In the Italian university system, salaries are established at the national level and fixed by academic rank and seniority. Thus all professors of the same academic rank and seniority receive the same salary, regardless of the university that employs them. The information on individual salaries is unavailable but the salaries ranges for rank and seniority are published. Thus we have approximated the salary for each individual as the average of their academic rank.
If information on salary is not available at all, one should at least compare research performance of individuals of the same academic rank.
The productivity of each scientist is calculated in each SDS and expressed on a percentile scale of 0-100 (worst to best) for comparison with the performance of all Italian colleagues of the same SDS; or as the ratio to the average performance of all Italian colleagues of the same SDS with productivity above zero[148]. In general we can exclude, for the Italian case, that productivity ranking lists may be distorted by variable returns to scale, due to different sizes of universities (Abramo, D'Angelo & Cicero, 2012d) or by returns to scope of research fields (Abramo, D'Angelo & Di Costa, 2012e).

[146] A preceding article by the same authors demonstrated that the average of the distribution of citations received for all cited publications of the same year and subject category is the most reliable scaling factor (Abramo, D'Angelo & Cicero, 2012c).
[147] The weighting values were assigned following advice from senior Italian professors in the life sciences. The values could be changed to suit different practices in other national contexts.
[148] In a preceding article the authors demonstrated that the average of the productivity distribution of researchers with productivity above 0 is the most effective scaling factor to compare the performance of researchers of different fields (Abramo, D'Angelo & Cicero, 2012f).

Labor productivity in a specific field

At field level, the yearly average productivity FSS_S over a certain period for researchers in a university (region, country, etc.) in a particular SDS[149] is:

$$FSS_S = \frac{1}{S_{RS}} \sum_{i=1}^{N} \frac{c_i}{\bar{c}_i} f_i \qquad [2]$$

Where:

S_{RS} = total salary of the research staff of the university in the SDS, in the observed period;

N = number of publications of the research staff in the SDS of the university, in the period of observation;

c_i = citations received by publication i;

$\bar{c}_i$= average citations received by all cited publications of the same year and subject category of publication i;

f_i = fractional contribution of researchers in the SDS of the university, to publication i, calculated as described above.

For each SDS we can construct a university (region, country, etc.) productivity ranking list by FSS_S expressed in percentiles or as the FSS_S ratio to average FSS_S of all universities with productivity above zero in the SDS.

The measures of productivity at field level permit identification of field strengths and weaknesses and thus correctly inform research policies and strategies.

Labor productivity of multi-fields units

In multi-field organizational units (i.e. disciplines, departments, universities, regions, nations), where there are researchers that belong to different fields, we are presented with the problem of how to aggregate productivity measures for researchers from the various fields. Two methods are possible, based on either the performance of individual researchers (FSS_R), or of the SDSs (FSS_S) present in the unit under examination. The appropriate choice depends on the objective for the measure. The first method emphasizes individual performance and the second emphasizes field performance, which we note is a "virtual" unit, since the members of the SDS at a university do not necessarily work together on a structured basis. The research administrator will perhaps be more interested in the performance results derived under the first method, determined from the average

[149] We note again that a field is not an organizational unit, rather a classification of researchers by their scientific qualifications. This does not mean that all the researchers in the same field and organization will necessarily form a single research group that works together. As an example, we quote the SDS description for FIS/03-Materials physics: "The sector includes the competencies necessary for dealing with theory and experimentation in the state of atomic and molecular aggregates, as well as competencies suited to dealing with properties of propagation and interaction of photons in fields and with material. Competencies in this sector also concern research in fields of atomic and molecular physics, liquid and solid states, semiconductors and metallic element composites, dilute and plasma states, as well as photonics, optics, optical electronics and quantum electronics". In the Italian academic system it is quite common to find "Materials physics" researchers working in two different departments (physics and engineering) at the same university.

of individual productivities. On the other hand the policy-maker, not being particularly interested in the performance variability within the organizational units but rather in comparison of the overall productivity of the various research institutions, could prefer the performance measure calculated by the second method. In the following subsections we present the two measurement procedures.

Labor productivity of multi-fields units based on FSS_R

We have seen that the performance of the individual researchers in a unit can be expressed in percentile rank or standardized to the field average. Thus the productivity of multi-field units can be expressed by the simple average of the percentile ranks of the researchers. It should be noted that the resort to percentile rank for the performance measure in multi-filed units or for simple comparison of performance for researchers in different fields is subject to obvious limitations, the first being compression of the performance differences between one position and the next. Thompson (1993) warns that *percentile ranks should not be added or averaged, because percentile is a numeral that does not represent* equal-interval measurement. Further, percentile rank is also sensitive to the size of the fields and to the performance distribution. For example, consider a unit composed of two researchers in two different SDSs (A and B, each with a national total of 10 researchers), who both rank in third place, but both with productivity only slightly below that of the first-ranked researchers in their respective SDSs: the average rank percentile for the unit will be 70. Then consider another unit with two researchers belonging to another two SDSs (C and D, each with 100 researchers), where both of the individuals place third but now with a greater gap to the top scientists of their SDSs (potentially much greater): their percentile rank will be 97. In this particular example, a comparison of the two units using percentile rank would certainly penalize the former unit.

However the second approach, involving standardization of productivity by field average, takes account of the extent of difference between productivities of the individuals. In formula, the productivity FSS_D over a certain period for department *D*, composed of researchers that belong to different SDSs:

$$FSS_D = \frac{1}{RS}\sum_{i=1}^{RS}\frac{FSS_{R_i}}{\overline{FSS_{R_\iota}}} \qquad [3]$$

Where:

RS = research staff of the department, in the observed period;

FSS_{R_i}= productivity of researcher *i* in the department;

$\overline{FSS_{R_\iota}}$= average productivity of all productive researchers in the same SDS of researcher *i*.

Labor productivity of multi-fields units based on FSS_S

The second method for measurement of research unit productivity involves identifying all the SDSs present in the unit and assigning each one a relative

weight depending on size (full time equivalent research personnel). As an example, for measurement of productivity of a university (region, nation) in a discipline (UDA), beginning from the productivity of the individual SDSs (FSSs), the productivity FSS_U of a university in a specific UDA *U*, is:

$$FSS_U = \sum_{i=1}^{N} \frac{FSS_{S_i}}{\overline{FSS_{S_i}}} \frac{S_{RS_i}}{S_{RS_u}} \quad [4]$$

With:
S_{RS_i} = total salary of the research staff of the university in the SDS *i*, in the observed period;
S_{RS_u} = total salary of the research staff of the university in the UDA *U*, in the observed period;
N= number of SDSs of the university in the UDA *U*;
$\overline{FSS_{S_i}}$ = national FSS_S in the SDS *i*.

For the measure of the productivity of a department (or university, region, country), the procedure is exactly the same: the only thing that changes is the size weight of the SDS, which is no longer with respect to the other SDSs of the UDA, but rather to all the SDSs of the department (university, region, country).
As noted, the appropriate choice between the two methods of measure for performance of a multi-field unit depends on the aims of the evaluation. The first method, based on productivity of individual researchers, interprets the performance of the unit as the average of the individual performances, meaning that the emphasis is on the individual. The other method, based on productivity of fields, interprets the field as a unique group (even though a virtual group), meaning that emphasis is on the overall product of the researchers that belong to the field, independently of the variability of the individual contributions. The two methods lead to performance results that are quite similar. In a future work we will provide a comparative in-depth analysis of the two methods.

Conclusions and recommendation

Until now, bibliometrics has proposed indicators and methods for measuring research performance that are largely inappropriate from a microeconomics perspective. The h-index and most of its variants, for example, inevitably ignore the impact of works with a number of citations below h and all citations above h of the h-core works. The h-index fails to field-normalize citations, to account for the number of co-authors and their order in the byline, Last but not least, because of the different intensity of publications across fields, productivity rankings need to be carried out by field (Abramo & D'Angelo, 2007), when in reality there is a human tendency to compare h-indexes for researchers across different fields. Each one of the proposed h-variant indicators tackles one of the many drawbacks of the h-index while leaving the others unsolved, so none can be considered completely satisfactory.

The new crown indicator, on the other hand, measures the average standardized citations of a set of publications, which cannot provide any indication of unit productivity. In fact a unit with double the MNCS value of another unit could actually have half the productivity, if the second unit produced four times as many publications. Whatever the CWTS research group (Waltman et al., 2012) might claim for them, the annual world university rankings by MNCS are not "performance" rankings - unless someone abnormally views performance as average impact of product, rather than impact per unit of cost. Applying the CWTS method, a unit that produces only one article with 10 citations has better performance than a unit producing 100, where each but one of these gets 10 citations and the last one gets nine citations. Further, the methodology reported for producing the ranking lists does not describe any weighting for co-authorship on the basis of byline order. Similar drawbacks are embedded is the SCImago Institutions Ranking by their main indicator, the Normalized Impact, measuring the ratio between the average scientific impact of an institution and the world average impact of publications of the same time frame, document type and subject area. We do not further consider any of the many annual world institutional rankings that are severely size dependent: the SJTU Shanghai Jiao Tong University, THES Times Higher Education Supplement and QS Quacquarelli Symonds rankings, among others. These seem to represent skilled communications and marketing operations, with the actual rankings resulting more from improvisation than scientifically-reasoned indicators and methods.

The great majority of the bibliometric indicators and the rankings based on their use present two fundamental limits: lack of normalization of the output value to the input value, and absence of classification of scientists by field of research. Without normalization there cannot be any measure of productivity, which is the quintessential indicator of performance in any production unit; without providing field classification of scientists, the rankings of multi-field research units will inevitably be distorted, due to the different intensity of publication across fields. An immediate corollary is that it is impossible to correctly compare productivity at international levels. In fact there is no international standard for classification of scientists and, we are further unaware of any nations that classify their scientists by field at domestic level, apart from Italy. This obstacle can in part be overcome by indirectly classifying researchers according to the classification of their scientific production into WoS or Scopus categories, and then identifying the predominant category. Fractional Scientific Strength (FSS) is a proxy indicator of productivity permitting measurement at different organizational levels. Both the indicator and the related methods can certainly be improved, however they do make sense according to economic theory of production. Other indicators and related rankings, such as the simple number (or fractional counting) of publications per research unit, or the average normalized impact, cannot alone provide evaluation of performance - however they could assume meaning if associated with a true measure of productivity. In fact if a research unit achieves average levels of productivity this could result from average production and

average impact, but also from high production and low impact, or the inverse. In this case, knowing the performance in terms of number of publications and average normalized impact would provide useful information on which aspect (quantity or impact) of scientific production to strengthen for betterment of production efficiency.

Aside from having an indicator of research unit productivity, the decision-maker could also find others useful, such as ones informing on unproductive researchers, on top researchers (10%, 5%, 1%, etc.), top publications, dispersion of performance within and between research units, etc.

Based on the analyses above, we issue an appeal and recommendation. Our appeal to scholars is to concentrate their efforts on the formulation of productivity indicators more or less resembling the one we propose, and on the relative methods of measurement, aiming at truly robust and meaningful international comparisons. Our recommendation is to avoid producing research performance rankings by invalid indicators and methods, which under the best of circumstances serve no effective purpose, and when used to inform policy and administrative decisions can actually be dangerous. Our undertaking, as soon as possible, should be to develop a roadmap of actions that will achieve international performance rankings that are meaningful and useful to the research administrator and policy-maker.

References

Abramo, G., Cicero, T. & D'Angelo, C.A. (2012a). A sensitivity analysis of researchers' productivity rankings to the time of citation observation. *Journal of Informetrics*, 6(2), 192–201.

Abramo, G., Cicero, T. & D'Angelo, C.A. (2012b). What is the appropriate length of the publication period over which to assess research performance? *Scientometrics*, 93(3), 1005-1017.

Abramo, G., Cicero, T. & D'Angelo, C.A. (2012c). Revisiting the scaling of citations for research assessment. *Journal of Informetrics*, 6(4), 470–479.

Abramo, G., Cicero, T. & D'Angelo, C.A. (2012d). Revisiting size effects in higher education research productivity. *Higher Education*, 63(6), 701-717.

Abramo, G., D'Angelo, C.A. & Di Costa, F. (2012e). Investigating returns to scope of research fields in universities. *Working paper LabRTT.* A short abstract available at http://www.disp.uniroma2.it/laboratoriortt/TESTI/Working%20paper/Returns_to_scope.pdf

Abramo, G., Cicero, T. & D'Angelo, C.A. (2013). Individual research performance: a proposal for comparing apples to oranges. *Journal of Informetrics*, 7(2), 528-529.

Abramo, G., D'Angelo, C.A. & Di Costa, F. (2011). Research productivity: are higher academic ranks more productive than lower ones? *Scientometrics*, 88(3), 915-928.

Abramo, G., D'Angelo, C.A. & Di Costa, F. (2008). Assessment of sectoral aggregation distortion in research productivity measurements. *Research Evaluation*, 17(2), 111-121.

Abramo, G. & D'Angelo, C.A. (2007). Measuring science: irresistible temptations, easy shortcuts and dangerous consequences. *Current Science*, 93(6), 762-766.

Archambault, É., Campbell, D., Gingras, Y. & Larivière, V. (2009). Comparing bibliometric statistics obtained from the Web of Science and Scopus. *Journal of the American Society for Information Science and Technology*, 60(7), 1320-1326.

Butler, L. (2007). Assessing university research: A plea for a balanced approach. *Science and Public Policy*, 34(8), 565-574.

Garfield, E. (1979). Is citation analysis a legitimate evaluation tool? *Scientometrics,*1(4), 359-375.

Glänzel, W. (2008). Seven myths in bibliometrics. About facts and fiction in quantitative science studies. Kretschmer & F. Havemann (Eds): *Proceedings of WIS Fourth International Conference on Webometrics, Informetrics and Scientometrics & Ninth COLLNET Meeting*, Berlin, Germany.

Hirsch, J.E. (2005). An index to quantify an individual's scientific research output. *Proceedings of the National Academy of Sciences*, 102(46), 16569-16572.

Leydesdorff, L. & Opthof, T. (2011) Remaining problems with the "New Crown Indicator" (MNCS) of the CWTS, *Journal of Informetrics*, 5(1), 224-225.

Moed, H.F. (2005). *Citation Analysis in Research Evaluation*. Springer, ISBN: 978-1-4020-3713-9

Moed, H.F., Burger, W.J M., Frankfort, J.G. & Van Raan, A.F.J. (1985). The application of bibliometric indicators: Important field- and time-dependent factors to be considered. *Scientometrics*, 8(3-4), 177-203.

Opthof, T. & Leydesdorff, L. (2010). Caveats for the journal and field normalizations in the CWTS ("Leiden") evaluations of research performance. *Journal of Informetrics*, 4(3), 423–430.

Pontille, D. (2004). *La Signature Scientifique: Une Sociologie Pragmatique de l'Attribution*. CNRS ÉDITIONS, Paris, 2004.

QS-Quacquarelli Symonds (2011). *World University Rankings*. Retrieved January 16, 2013 from: http://www.topuniversities.com/university-rankings/world-university-rankings.

RIN (Research Information Network) (2009). *Communicating Knowledge: How and Why Researchers Publish and Disseminate Their Findings*. London, UK: RIN. Retrieved January 16, 2013 from: http://www.jisc.ac.uk/publications/research/2009/communicatingknowledgereport.aspx.

SJTU-Shanghai Jiao Tong University (2011). *Academic Ranking of World Universities.* Retrieved January 16, 2013 from: http://www.shanghairanking.com/ARWU2011.html.

THES-Times Higher Education Supplement (2012). *World Academic Ranking 2011-2012*. Retrieved January 16, 2013 from: http://www.timeshighereducation.co.uk/world-university-rankings/2011-2012/top-400.html.

Thompson, B. (1993). GRE Percentile Ranks Cannot Be Added or Averaged: A Position Paper Exploring the Scaling Characteristics of Percentile Ranks, and the Ethical and Legal Culpabilities Created by Adding Percentile Ranks in Making "High-Stakes" Admission Decisions. Paper presented at the *Annual Meeting of the Mid-South Educational Research Association*, New Orleans, LA, November 12, 1993.

Waltman, L., Van Eck, N.J., Van Leeuwen, T.N., Visser, M.S. & Van Raan, A.F.J. (2011). Towards a new crown indicator: Some theoretical considerations. *Journal of Informetrics*, 5(1), 37–47.

Waltman, L., Calero-Medina, C., Kosten, J., Noyons, E. C. M., Tijssen, R. J. W., Van Eck, N. J., ... & Wouters, P. (2012). The leiden ranking 2011/2012: Data collection, indicators, and interpretation. *Journal of the American Society for Information Science and Technology*, 63(12), 2419-243.

THE ROLE OF NATIONAL UNIVERSITY RANKINGS IN AN INTERNATIONAL CONTEXT: THE CASE OF THE I-UGR RANKINGS OF SPANISH UNIVERSITIES

Nicolás Robinson-García[1], Jose Garcia Moreno-Torres[2], Daniel Torres-Salinas[3], Emilio Delgado López-Cózar[1] and Francisco Herrera[4]

[1] *{elrobin, edelgado}@ugr.es*
EC3: Evaluación de la Ciencia y de la Comunicación Científica, Departamento de Biblioteconomía y Documentación, Universidad de Granada (Spain)

[2] *jose.garcia.mt@gmail.com*
Department of Computer Science and Artificial Intelligence, Universidad de Granada (Spain)

[3] *torressalinas@gmail.com*
EC3: Evaluación de la Ciencia y de la Comunicación Científica, Centro de Investigación Biomédica Aplicada, Universidad de Navarra (Spain)

4 herrera@decsai.ugr.es
Department of Computer Science and Artificial Intelligence, Universidad de Granada (Spain)

Abstract

The great importance international rankings have in the research policy arena calls for caution as they present many flaws and shortcomings. One of them has to do with the inability to accurately represent national university systems as their original purpose is only to rank world-class universities. Another one has to do with the lack of representativeness of universities' disciplinary profiles as they usually provide a unique table. Although some rankings offer a great coverage and others offer league tables by fields, no international ranking does both. In order to surpass such limitation from a research policy viewpoint, this paper analyzes the possibility of using national rankings in order to complement international rankings. For this, we describe the Spanish university system as a study case presenting the I-UGR Rankings for Spanish universities by fields and subfields. Then, we compare their results with those obtained by the Shanghai Ranking, the QS Ranking and the NTU Ranking, as they all have basic common grounds which allow such comparison. We conclude that it is advisable to use national rankings in order to complement international rankings, however we observe that this must be done with certain caution as they differ on the methodology employed as well as on the construction of the fields.

Conference Topic

Science Policy and Research Evaluation: Quantitative and Qualitative Approaches (Topic 3) and Management and Measurement of Bibliometric Data within Scientific Organizations (Topic 9).

1. Introduction

Since the launch of the first edition of the Shanghai Ranking in 2003, interest has grown on the development of tools for benchmarking and comparing academic and research institutions. As a result of the massification of higher education, the race for excellence and a fierce battle for research funding, universities now strive for positioning themselves in these international rankings (Hazelkorn 2011). These tools have gain an undisputable position in the research managers 'toolkit' for measuring the state of health of higher education institutions and the main resource for many universities and countries when taking decisions in a research policy context (Marginson & van der Wende, 2007). The great effect they have, - not only in the media and the public but also for research managers, politicians and decision makers, - relies on the perception that highly ranked institutions are usually more productive, produce higher quality research and teaching and contribute best to society than the rest of universities (Shin & Toutkoushian, 2011).

However, despite their advantages as easy-to-read tools, they also have many inconsistencies and shortcomings that warn against a careless use (Delgado López-Cózar, 2012). In this sense, we can identify five major issues which must be addressed: 1) methodological and technical errors and difficulties such as the recollection of reliable and standardized data (Toutkoushian & Webber, 2011), 2) the criteria for selecting the indicators are not scientifically supported (Van Raan, 2005), 3) the multidimensional nature of universities (Orduña-Malea, 2012; Waltman et al., 2012) leads to a wide heterogeneity among institutions (Collini 2011), 4) using a unique table to rank universities neglects their disciplinary focus (Visser et al., 2007), and 5) international rankings cannot reflect the state of national higher education systems as they usually cover just the top universities of each country (Torres-Salinas et al., 2011a).

While the issue of data reliability still remains a major shortcoming and there is no consensus yet over which indicators represent better the nature and quality of universities, the other issues have been somehow surpassed using approaches which do not solve completely their dangers but at least, diminishes the flaws. For instance, rankings such as the Leiden Ranking (Waltman et al., 2012) or the Scimago Institutions Rankings (henceforth SIR) have emerged focusing uniquely on the research dimension of universities to the neglect of other aspects such as innovation, transference of knowledge or teaching. Others, such as the Shanghai Ranking, the Times Higher Education World University Rankings (henceforth THE Ranking), the QS Rankings or the National Taiwan University Ranking

(henceforth NTU Ranking, previously produced by the Higher Education Evaluation and Accreditation Council of Taiwan) now publish, along with a global ranking, rankings by subjects and fields, which offer a better picture of universities' performance (García et al., 2012). Also, some rankings such as the SIR or the Ranking Web of World Universities cover now not just top-class universities but the former includes more than 3,000 research institutions and the latter, more than 19,000.

Rankings have not been fully developed and still draw serious shortcomings (van Raan, 2005). But their dominance as decisive factors in research policy (Hazelkorn, 2011) at national and supranational level puts them in the spotlight. One of the most important threats rankings entail is that they ignore universities' diversity, which can affect seriously the health of higher education systems and lead to dangerous and simplistic conclusions when interpreting and developing ranking systems (e.g., Moed et al., 2011). These differences affect institutions at two levels, at their organizational structure, and in the national configuration of higher education systems, affecting their multidisciplinary nature and diversity (Orduña, 2012). The phenomenon of university rankings has influenced deeply all university systems, even those that were not conceived at first to establish a competitive framework. Therefore, in order to analyze the success or failure of different countries in their research policy, university systems should be assessed as a whole, and not considering each university as an individual and autonomous unit. Such approach was applied by Docampo (2011) using the Shanghai Ranking in order to analyze the university systems of the countries represented.

Despite its limitations, this study offers a glimpse of the global scenario regarding the research excellence of different countries' university systems. In Table 1 we show the clusters emerged from the study carried out by Docampo (2011) and the number of universities by country in different intervals according to the 2012 edition of the Shanghai Ranking. Therefore we observe a dominance of the United States and the United Kingdom which alone, represent more than a third of the universities included in the ranking (37.6%), followed by Germany and Canada as the next with the highest number of universities included. However, despite the numbers, except Japan, which in this new edition includes a university in the top20, none of the others have a university positioned within this interval. In this context, the truth is that the high visibility Anglo-Saxon universities have in rankings leaves little space for others, blurring the state of other countries which are working towards a successful university model. In fact, it clearly shows the incapability of the ranking to represent national university systems with exhaustiveness.

Thus, these rankings do not offer a complete view of national higher education systems, preventing research managers and decision makers to have an accurate picture of the state of each country's university system. For this reason, in 2010

we developed the Rankings I-UGR of Spanish Universities according to Fields and Scientific Disciplines[150] (henceforth I-UGR Rankings) available at http://rankinguniversidades.es. This website offers 49 rankings for Spanish universities divided in 12 fields and 37 disciplines, according to their international research performance. Spain is a good example of a misrepresented higher education system. For instance, in the 2012 edition of the Shanghai Ranking only 11 universities out of 74 met the criteria for inclusion in the global ranking. In fact, none made it to the top 100 and only three were included in the 201-300 interval. Also, as it occurs with other countries such as Italy (Abramo, Cicero & D'Angelo, 2011), it is a non-competitive higher education system, which means that universities do not act as individual units but within a national framework, therefore decisions should not be taken relying in such a poor sample.

Table 1. University systems by country considering the results in Docampo (2011) and the 2012 Shanghai Ranking edition. Leaders, Fast followers and followers

	Countries	Nr of Universities Top20	Nr of Universities Top100	Nr of Universities Top300	Nr of Universities Top500
Leaders	United States	17	53	109	150
	United Kingdom	2	9	30	38
	Switzerland	---	4	7	7
Fast followers	Australia	---	5	9	19
	Canada	---	4	17	22
	Sweden	---	3	7	11
	Israel	---	3	4	6
	Netherlands	---	2	10	13
	Denmark	---	2	4	4
Followers	Germany	---	4	24	37
	France	---	3	13	20
	Belgium	---	1	6	7
	Norway	---	1	3	4
	Finland	---	1	1	5

The main goal of the present paper is to justify that national rankings are necessary in order to complement international rankings. For this we will use the I-UGR Rankings analyzing:

1) Levels of agreement with international rankings: are the top Spanish universities the ones visible in international rankings?

2) Disciplinary concordance: do the different classifications by fields and subjects allow an analysis by areas?

[150] I-UGR stands for Institutions - University of Granada.

The paper is structured as follows. First we present the Spanish case analyzing its current state and we introduce the I-UGR Rankings, we contextualize its creation and we describe the methodology employed for their development. Next, we address the main issue of this paper: we compare the results of the main international rankings and the I-UGR Rankings for Spanish universities. To do so, we selected the Shanghai Ranking, the QS Ranking and the NTU Ranking. Finally, in Section 4 we resume our main findings and their consequences in a research policy scenario.

2. Spain as a case study: introduction to the I-UGR Rankings

The Spanish university system is formed by 74 universities: 48 public and 26 private. However in the 2012 edition of the Shanghai Ranking only 11 met the minimum requirements to be included. It is a country poorly represented in the main international rankings due to the scarce number of universities considered as World-Class universities. But the impact these rankings have in research policy threatens a good governance and sensible decision making as they do not offer a complete picture of the university system (Docampo, 2011). In fact, as observed in Table 2, only 20 universities (19 public and 1 private universities) are included in three of the most important rankings; that is, 27.03% of the whole system. For this reason, other tools are needed in order to complete this fragmented picture of the Spanish higher education scenario.

Table 2. Spanish universities represented in the 2012 edition of the Shanghai Ranking, the QS Ranking and the NTU Ranking

Position of Spanish Universities in Shanghai Ranking		Position of Spanish Universities in QS Ranking		Position of Spanish Universities in NTU Ranking	
Barcelona	201-300	Autónoma de Barcelona	176	Barcelona	115
Autónoma de Madrid	201-300	Barcelona	187	Autónoma de Barcelona	191
Complutense de Madrid	201-300	Autónoma de Madrid	206	Autónoma de Madrid	231
Valencia	301-400	Complutense de Madrid	226	Valencia	253
Autónoma de Barcelona	301-400	Pompeu Fabra	266	Complutense de Madrid	259
Politécnica de Valencia	301-400	Carlos III de Madrid	343	Santiago de Compostela	330
País Vasco	301-400	Politécnica de Cataluña	350	Granada	335
Granada	401-500	Navarra	359	Zaragoza	382
Pompeu Fabra	401-500	Politécnica de Valencia	401-450	Pompeu Fabra	408
Zaragoza	401-500	Politécnica de Madrid	401-450	País Vasco	420
Vigo	401-500	Granada	451-500	Oviedo	461
		Salamanca	451-500	Politécnica de Valencia	471
		Santiago de Compostela	451-500	Sevilla	483
		Valencia	451-500		
		Zaragoza	501-550		
		Sevilla	551-600		
		Alcalá de Henares	601+		
		Murcia	601+		

The first edition of the I-UGR Rankings was launched on 2010. Its development was motivated by the scarce visibility Spanish universities have in international rankings, which leads to a fragmented picture of the Spanish university system. Though other national rankings had already been developed, these were considered insufficient due to the limitations they presented which made them unsuitable as research policy tools. Among other limitations we address the following: lack of continuity over time, exclusion of private institutions, disregard of disciplinary focus, use of rudimentary bibliometric indicators, selection of unsuitable time periods or election of databases with dubious selection criteria of sources (Torres-Salinas et al., 2011a).

Data is retrieved from the Thomson Reuters Web of Science database. In its first edition 12 rankings were offered for 12 broad fields. These fields were later expanded with 19 subfields or disciplines in the second edition (Torres-Salinas et al., 2011b) and finally, 37 disciplines in the 2012 edition. The fields and disciplines were constructed by aggregating the subject categories to which records from the Science Citation Index and Social Science Citation Index are assigned. Aggregating subject categories is a classical perspective followed in many bibliometric studies when adopting a macro-level approach (e.g., Moed, 2005; Leydesdorff & Rafols, 2009). For further information on the coverage on the I-UGR Rankings and the development of the fields and subfields the reader is referred to the Methodology section of the rankings' website available at http://rankinguniversidades.es. Once the data is compiled into a relational database, the indicators defined in Table 3 are computed and normalized in [0, 1], and the index for rating each university is calculated. To rank universities we use the IFQ^2A Index (Torres-Salinas et al., 2011c). This indicator measures the quantitative and qualitative dimensions of the research outcome of a group of institutions in a given field. It is based on six primary bibliometric indicators, three focused on the quantitative dimension (QNIF) and the other three focused on the qualitative dimension (QLIF). In Table 3 we summarize the methodology employed for calculating the IFQ^2A Index. For a detailed explanation on the IFQ^2A Index the reader is referred to Torres-Salinas et al. (2011c).

Table 3. Calculation of the IFQ2A Index and definition of indicators.

$QNIF = \sqrt[3]{NDOC \times NCIT \times H}$		$QLIF = \sqrt[3]{\%1Q \times ACIT \times TOPCIT}$	
NDOC	Number of citable papers published in scientific journals	%1Q	Ratio of papers published in journals in the top JCR quartile
NCIT	Number of citations received by all citable papers	ACIT	Average number of citations received by all citable papers
H	H-Index as proposed by Hirsch (2005), over all the publications of the institution	TOPCIT	Ratio of papers belonging to the top 10% most cited papers calculated within all institutions
$IFQ^2A = QNIF \times QLIF$			

The selection of the indicators as well as the conceptualization of the index, are based on the following criteria:

1) The indicators chosen must not be restrictive. That is, they should be applied to all institutions. For instance, the Shanghai Ranking uses the number of Nobel Prizes as an indicator to measure research excellence. In the Spanish case only one university is affected by it (Complutense de Madrid).

2) Rankings must be size-independent. This leads to the use of a bidimensional index which takes into account research outcome but also excellence, benefiting equally: small and large institutions.

3) Rankings must take into account the disciplinary focus of universities. For this, a unique list cannot be provided. Contrarily one most offer rankings by field of specialization in order to provide useful tools for research managers.

4) Seniority must not be rewarded. For this fixed time periods must be used. Also, when calculating the H-Index, this must be considering the time frame used. In this sense, the I-UGR Rankings offer a five-year window and a ten-year window.

5) Stability must be assured. This means that the fixed time frame must be wide enough to offer stable results. A five-year time frame allows results to be consistent and significant.

In Figure 1 we show the distribution of universities according to the QNIF and QLIF in the field of Medicine and Pharmacology for the 2007-2011 time period. The dashed lines show the average values of each dimension. Universities positioned at the top right hand of the figure are those which outstand in both dimensions. Those positioned on the bottom right outstand on the quantitative dimension but not on the qualitative dimension. At the top left, we observe university with small research output but high quality research. Lastly, in the bottom left, universities which do not outstand in any dimension are represented. As we can observe, although top universities outstand in both dimensions, many universities outstand in the qualitative dimension but do not do so in the quantitative dimension. Due to the bidimensional nature of the IFQ^2A index, these small institutions are reflected in the rankings.

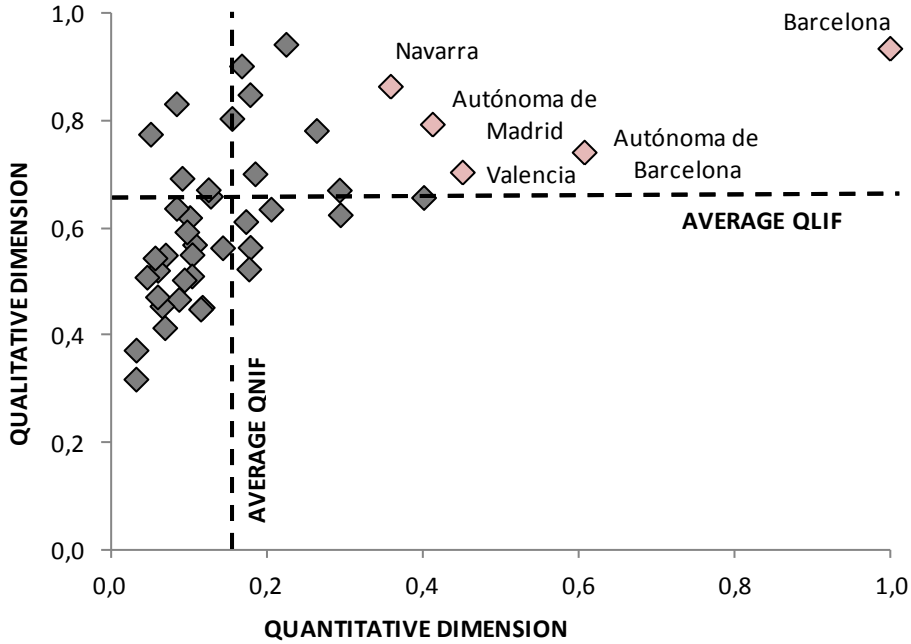


Figure 1. Distribution of universities according to their qualitative and quantitative dimensions in the field of Medicine and Pharmacology. 2007-2011.

3. Comparison by fields of the main international rankings and the I-UGR for Spanish universities

In this section we analyze the state of the Spanish university system using international and national rankings. For this, we first establish in Section 3.1 a set of criteria for the selection of the rankings we will use in order to set some basic common grounds which will allow a fair comparison between them. Then, in Section 3.2 we match rankings by fields between the international and national rankings and finally, we analyze the level of agreement between them. For this we use two indicators. On the one hand, we calculate the Spearman's rank correlation coefficient or Spearman's rho, which will indicate to what extent are the different rankings coherent between them. On the other hand, we show the level of agreement between rankings, which indicates if universities included in an international ranking coincide with those which occupy the top positions of the national ranking.

3.1 Selection of rankings

The aim is to use international and national rankings as complementary tools to offer on the one hand, a global perspective of the position of Spanish universities and, on the other hand, a complete picture of the Spanish university system. For

this, we first need to establish a set of criteria for choosing the most relevant rankings for our purposes. These are the following:

1) As we are analyzing the research dimension of universities, rankings must be based on the research performance of universities, at least partially.

2) Data retrieved for the construction of the rankings must come from a reliable bibliometric database or information resource, at least partially.

3) They must offer rankings by fields, as we have considered that only this way we can provide an accurate image of universities' research performance.

Based on these criteria we selected the I-UGR Rankings as national rankings and the following international rankings. The methodology of each ranking is available at its website, due to space limitations it has not been included in this paper:

1) *Shanghai Ranking* (http://www.shanghairanking.com/). It was not only the first international ranking launched (Liu & Cheng, 2005) but it is used as yardstick to measure the research excellence of universities worldwide (Docampo, 2011). It is based on six indicators, two of them (40% of the total rating) are based on data retrieved from the Web of Science (for more information on this ranking the reader is referred to Liu & Cheng, 2005; van Raan, 2005; Docampo 2011; Aguillo et al., 2010). Since 2007 it offers five rankings by field and since 2009, five ranking by subject.

2) *QS Ranking* (http://www.topuniversities.com/). The first edition of this ranking was launched in 2004. Until 2009 it was produced in partnership with the Times Higher Education, however, since then each company develops its own ranking (for more information on this ranking the reader is referred to Aguillo et al., 2010; Usher & Savino, 2007). 20% of the total rating assigned to each university is based on data retrieved from the database Scopus. It offers along with the global league table, 29 rankings by discipline classified into five major fields.

3) *NTU Ranking* (http://nturanking.lis.ntu.edu.tw). This ranking was first launched in 2007. It aims at measuring solely the quality of universities' research. It is based on 8 indicators all of them supported by bibliometric data from the Web of Science and the Thomson Reuters Essential Science Indicators (for more information on this ranking the reader is referred to e.g., Aguillo et al., 2010). Along with the global table league, it offers rankings by field and subject in a similar structure to that of the Shanghai Ranking. In this case, it offers 6 rankings by field and 14 rankings by subject.

3.2 Concordance between international and national rankings and levels of agreement

In order to establish fair comparisons and provide a global picture of the state of Spanish universities using national and international rankings, we first need to ensure that the classification of fields of national and international rankings is somehow similar and therefore, compatible. For this, we would need to analyze the way these fields are constructed for the four rankings used in this study and determine to which grade the methodology employed by each of them allows fair comparisons. As mentioned before, the I-UGR Rankings construct fields and disciplines by aggregating the Thomson Reuters subject categories. The NTU Ranking uses the same approach, and the construction of fields and subjects is declared at their website (http://nturanking.lis.ntu.edu.tw). However, this does not occur for the other two rankings, which do not declare the methodology employed for establishing such fields. This lack of transparency is a shortcoming that must be taken into account when using these rankings for research policy.

Table 4. Matching of fields and disciplines between the Shanghai Ranking and the I-UGR Rankings

SHANGHAI RANKING	I-UGR RANKINGS	RHO	A
Natural Sciences & Mathematics	Mathematics / Physics / Chemistry	-0,50; -0,50; 0,50	0/3; 3/3; 2/3
Engineering/Technology & Computer Sciences	Engineering / Information & Communication Technology	*	1/3; 2/3
Life & Agricultural Sciences	Agricultural Sciences / Biological Sciences	1,00; 1,00	1/2; 2/2
Clinical Medicine & Pharmacy	Medicine & Pharmacy	1,00	2/2
Social Science	Other Social Sciences / Psychology & Education / Economics, Finance & Business	*	0/2; 0/2; 2/2
Mathematics	Mathematics	-0,23	4/8
Physics	Physics	0,72	5/5
Chemistry	Chemistry	0,26	8/10
Computer Science	Computer Science	0,41	3/6
Economics & Business	Economics, Finance & Business	*	2/2

Note: Rho indicates the Spearman's coefficient. A indicates the level of agreement between rankings, that is, the number of universities present in both rankings.
*Insufficient values to calculate the indicator

We analyzed the fields and subjects of the selected international rankings and we established the homologous field or discipline according to the I-UGR Rankings. In Tables 4-6 we show the matching of fields per ranking. In general terms, we observe that it is possible to match most of the fields between the three international rankings selected and the I-UGR Rankings, although some exceptions are noted. The areas misrepresented in the I-UGR Rankings were Mechanical Engineering (QS Ranking and NTU Ranking), Law (QS Ranking) and all of the areas considered of the Arts & Humanities fields by the QS

Ranking. This is due to the way the I-UGR Rankings are constructed, as they rely on the JCR and these lack journal rankings for these fields. Also, we observe that some fields of the international rankings (i.e., the Shanghai Ranking and the field of Social Science) include more than one of the tables by field of the I-UGR Rankings. Finally, the classification of fields and subfields does not always match between rankings. Although this issue has no relevance for the purposes of this analysis, we must point out that subjects considered as major areas in one ranking are considered in the other as subfields or disciplines.

Table 5. Matching of fields and disciplines between the QS Ranking and the I-UGR Rankings

	QS RANKING	I-UGR RANKINGS	RHO	A
Arts & Humanities	Philosophy			
	Modern Languages			
	Geography	Geography & City Planning	0,68	2/6
	History			
	Linguistics			
	English Language & Literature			
Engineering & Technology	Computer Science & Information Systems	Computer Science	-0,87	1/3
	Chemical Engineering	Chemical Engineering	0,84	4/7
	Civil Engineering	Civil Engineering	-0,5	1/3
	Electrical Engineering	Electric & Electronic Engineering	-0,43	4/7
	Mechanical Engineering			
Life Sciences & Medicine	Medicine	Medicine	0,50	3/3
	Biological Sciences	Biological Sciences	0,87	3/3
	Psychology	Psychology	0,26	6/7
	Pharmacy & Pharmacology	Pharmacy & Toxicology	0,74	3/5
Natural Sciences	Physics & Astronomy	Physics	0,67	4/5
	Mathematics	Mathematics	0,21	2/4
	Environmental Sciences	Earth & Environmental Sciences	0,87	2/3
	Earth & Marine Sciences	Earth & Environmental Sciences	1,00	1/2
	Chemistry	Chemistry	0,80	3/4
	Materials Science	Materials Science	0,83	3/6
Social Sciences & Management	Statistics & Operational Research	Statistics	-0,62	3/6
	Sociology	Sociology	-1,00	1/2
	Politics & International Studies	Political Science	**	0/1
	Law			
	Economics & Econometrics	Economics	0,50	4/6
	Account & Finance	Business	0,87	2/3
	Communication & Media	Communication	0,00	0/3
	Education	Education	0,29	1/5

Note**:** Rho indicates the Spearman's coefficient. A indicates the level of agreement, that is, the number of universities present in both rankings.
*Insufficient values to calculate the indicator

The three selected rankings included a total of 30 Spanish universities dispersed in 40 different fields and subfields. In Tables 4-6 we show the levels of agreement between international and national rankings according to the assignment of areas.

For each area we calculate the Spearman coefficient to analyze the consistency between both rankings and the number of universities included in international rankings which take up the top positions of the national ranking. That is, if 6 Spanish universities are included in an international ranking but only two occupy positions between 1 and 6, the coincidence will be 2/6.

Table 6. Matching of fields and disciplines between the NTU Ranking and the I-UGR Rankings

NTU RANKING	I-UGR RANKINGS	RHO	A
Agriculture	Agriculture	0,34	10/13
Clinical Medicine	Medicine	1,00	2/3
Engineering	Engineering	0,19	9/11
Life Sciences	Biological Sciences	0,77	6/6
Natural Sciences	Mathematics / Physics / Chemistry & Chemical Engineering	0,14; 0,94; 0,75	7/10; 8/10; 9/10
Social Sciences	Other Social Sciences / Psychology & Education / Economics…	0,36; -0,69; 0,95	3/4; 3/4; 2/4
Agricultural Sciences	Agricultural Sciences	0,38	17/21
Environment/Ecology	Earth & Environmental Sciences	0,53	6/9
Plant & Animal Science	Biological Sciences	0,55	6/10
Computer Science	Computer Science	0,754	8/13
Chemical Engineering	Chemical Engineering	0,55	8/11
Civil Engineering	Civil Engineering	0,47	8/12
Electrical Engineering	Electrical & Electronic Engineering	0,58	8/11
Mechanical Engineering			
Materials Science	Materials Science	1,00	4/5
Pharmacology & Toxicology	Pharmacy & Toxicology	0,6	5/5
Chemistry	Chemistry	0,84	14/15
Geosciences	Geosciences	0,89	5/6
Mathematics	Mathematics	0,88	11/12
Physics	Physics	0,89	6/7

Note**:** Rho indicates the Spearman's coefficient. A indicates the level of agreement. *Insufficient values to calculate the indicator

The Shanghai Ranking is the less consistent with the I-UGR Rankings as only two fields have significant correlations (Life & Agricultural Sciences and Physics), while the NTU Ranking shows high correlations in 11 out of 20 fields (Table 6) and the QS Ranking correlates in 7 out of 21(Table 5). The three fields with the highest correlations between the NTU Ranking and the I-UGR Rankings are Clinical Medicine (1,00), Materials Sciences (1,00) and Natural Sciences (0,94 with Physics). In the case of the QS Ranking, these three fields are Earth & Marine Sciences (1,00) and, Biological Sciences, Environmental Sciences and Account& Finance, all of them with a value of 0,87. The fields with high correlation belong in most cases to the fields of Biomedicine, Life Sciences and

Exact Sciences, and the ones with least correlation belong to the Social Sciences. Only one exception is noted in the field of Social Science for the NTU Ranking, which has a high correlation with the field of Economics of the I-UGR Rankings.

4. Conclusions

In this paper we explore the possibility of using national rankings to complement international rankings, as the latter usually offer a poor representation of national university systems. We insist on the importance of rankings by fields (García et al., 2012) as these do not neglect universities' disciplinary focus and offer a complete picture of universities' research performance. For this we use Spain as a study case and we introduce the I-UGR Rankings for Spanish universities. This ranking uses the IFQ^2A Index, an indicator which measures the qualitative as well as the quantitative dimension of research (Torres-Salinas, 2011c). Then, we select three international rankings (Shanghai Ranking, QS Ranking and NTU Ranking) according to a given set of criteria; we analyze the concordance between the fields these rankings offer and the ones given by the I-UGR Rankings in order to establish equivalences between them. Finally, we calculate the Spearman's coefficient and we analyze the levels of agreement between the universities included in the international rankings and the top positions of the national rankings. From this analysis we conclude that national rankings can complement international rankings in order to provide a complete picture of university systems despite the methodological differences aroused from the comparisons by fields.

Although there are differences between the methodologies employed by the various rankings, it is possible to use both and combine them in a research policy context. The coherence between them is especially significant for the fields of Biomedicine, Life Sciences and Exact Sciences. This does not occur in the Social Sciences where the only exception noted is Economics. In general terms, the NTU Ranking is the one which seems to be more consistent with the I-UGR Rankings. This is not surprising as it is the only one which measures solely the research dimension and fully based on the Web of Science, as it occurs with the I-UGR Rankings. Also, the confection of the fields and subfields is similar as both rankings aggregate subject categories to construct the fields, while in the other two cases this is not explained. Another issue which affects the correlation between rankings has to do with the way results are presented in the Shanghai Ranking and the QS Ranking, as they only show the intervals in which each university is positioned after they surpass certain threshold. Although the QS Ranking provides the rating of each university, allowing the user to rank universities, this those not occur with the Shanghai Ranking. Having said this and despite of the shortcomings mentioned, we observe coherent results between rankings leading us to assure that it is possible to use national rankings as complement to international rankings in order to offer a complete picture of national university systems in a research policy context.

Acknowledgments

Nicolás Robinson-García and Jose G. Moreno-Torres are currently supported by a FPU grant from the Ministerio de Educación y Ciencia of the Spanish Government.

References

Aguillo, I., Bar-Ilan, J., Levene, M. & Ortega, J.L. (2010). Comparing University Rankings. *Scientometrics*, 85(1), 243-256.

Abramo, G., Cicero, T. & D'Angelo, C.A. (2011). The Dangers of Performance - based Research Funding in Non-competitive Higher Education Systems. *Scientometrics*, 87(3), 641-654.

Aghion, P., Dewatripoint, M., Hoxby, C., Mas-Colell, A. & Sapir, A. (2010). The governance and performance of universities: Evidence from Europe and the US. *Economic Policy*, 25(61), 7-59.

Collini, S. (2011). *What are universities for?* London, UK: Penguin Books.

Delgado López-Cózar, E. (2012). Cómo se cocinan los rankings universitarios. *Dendra Médica. Revista de Humanidades*, 11(1), 43-58.

Dill, D.D. & Soo, M. (2005). Academic quality, league tables and public policy: a cross-national analysis of university ranking systems. *Higher Education*, 49(4),494-533.

Dobbins, M., Knill, C. & Vögtle, E.M. (2011). An Analytical Framework for the cross-country Comparison of Higher Education Governance. *Higher Education*, 62(5), 665-683.

Docampo, D. (2011). On the use of the Shanghai Ranking to assess the research performance of university systems. *Scientometrics*, 86(1), 77-92.

García, J.A., Rodríguez-Sánchez, R., Fdez-Valdivia, J., Robinson-García, N. & Torres-Salinas, D. (2012). Mapping Institutions According to their Journal Publication Profile: Spanish Universities as a Case Study. *Journal of the American Society for Information Science and Technology*, 63(11), 2328-2340.

Hazelkorn, E. (2011). *Rankings and the Reshaping of Higher Education. The Battle for World-Class Excellence*. Houndmills, UK: Palgrave MacMillan.

Hirsch, J.E. (2005). An index to quantify an individual's scientific research output. *Proceedings of the National Academy of Sciences USA*, 102(46), 16569–16572.

Labaree, D.F. Understanding the Rise of American Higher Education: How Complexity Breeds Autonomy. http://www.stanford.edu/~dlabaree/selected-papers.html.

Leydesdorff, L. & Rafols, I. (2009). A global map of science based on the ISI subject categories. *Journal of the Americna Society for Information Science and Technology*, 63(11), 2239-2253.

Liu, N.C. & Cheng, Y. (2005).The Academic Ranking of World Universities. *Higher Education in Europe*, 30(2), 127-136.

Marginson, S. & van der Wende, M. (2007). To Rank or to be Ranked: The Impact of Global Rankings in Higher Education. *Journal of Studies in International Education*, 11(3-4), 306-329.

Moed, H.F. (2005). *Citation Analysis in Research Evaluation.* Dordrecht: Springer.

Moed, H.F., Moya-Anegón, F., López-Illescas, C. & Visser, M. (2011). Is concentration of university research associated with better research performance? *Journal of Informetrics*, 5(4), 649-658.

Orduña-Malea, E. (2012). *Propuesta de un modelo de análisis redinformétrico multinivel para el estudio sistémico de las universidades* [Doctoral Thesis]. Valencia, Spain: Universidad Politécnica de Valencia.

The Research Universities Consortium (2012). *The Current Health and Future Well-being of the American Research University*. http://www.researchuniversitiesfutures.org/RIM_Report_Research%20Future's%20Consortium%20.pdf

Shin, J. C. & Toutkoushian, R.K. (2011). The Past, Present, and Future of University Rankings. In: Shin, J.C., Toutkoushian, R.K. & Teichler, U. (eds.). *University Rankings. Theoretical Basis, Methodology and Impacts on Global Higher Education* (pp. 1-18). Dordrecht, Netherlands: Springer.

Torres-Salinas, D., Delgado López-Cózar, E., Moreno-Torres, J.G. & Herrera, F. (2011a). Rankings ISI de las Universidades Españolas según Campos Científicos: Descripción y Resultados. *El Profesional de la Información*, 20(1), 111-118.

Torres-Salinas, D., Moreno-Torres, J.G., Robinson-García, N., Delgado López-Cózar, E. & Herrera, F. (2011b). Rankings ISI de las Universidades Españolas según Campos y Disciplinas Científicas (2ª ed. 2011). *El Profesional de la Información*, 20(6), 701-709.

Torres-Salinas, D., Moreno-Torres, J.G., Delgado López-Cózar, E. & Herrera, F. (2011c). A methodology for Institution-Field ranking based on a bidimensional analysis: the IFQ^{2}A index. *Scientometrics*, 88(3), 771-786.

Toutkoushian, R.K. & Webber, K. (2011). Measuring the Research Performance of Postsecondary Institutions. In: Shin, J.C., Toutkoushian, R.K. & Teichler, U. (eds.). *University Rankings. Theoretical Basis, Methodology and Impacts on Global Higher Education* (pp. 123-144). Dordrecht, Netherlands: Springer.

Usher, A. & Savino, M. (2007). A Global Survey of University Rankings and League Tables. *Higher Education in Europe*, 32(1), 5-15.

Van Raan, A.F.J. (2005). Fatal Attraction: Conceptual and Methodological Problems in the Ranking of Universities in the Global Scientific Communication System. *Scientometrics*, 62(1), 133-143.

Visser, M.S., Calero-Medina, C.M. & Moed, H.F. (2007). Beyond Rankings: The Role of Large Research Universities in the Global Scientific Communication System. In: Torres-Salinas, D. & Moed, H.F. *Proceedings of 11th Conference of the International Society for Scientometrics and Informetrics* (pp. 761-765). Madrid, Spain: CINDOC-CSIC.

Waltman, L., Calero-Medina, C., Kosten, J., Noyons, E.C.M., Tijssen, R.J., van Eck, N.J., et al. (2012). The Leiden Ranking 2011/2012: Data Collection, Indicators and Interpretation. In: Archambault, É., Gingras, Y. & Larivière, V. *Proceedings of 17th International Conference on Science and Technology Indicators* (791-802). Montreal, Canada: Science-Metrix and OST.

SCIENCE DYNAMICS: NORMALIZED GROWTH CURVES, SHARPE RATIOS, AND SCALING EXPONENTS

Haiko Lietz[1] and Mathias Riechert[2]

[1] *haiko.lietz@gesis.org*
iFQ – Institute for Research Information and Quality Assurance, Schützenstraße 6a, 10177 Berlin (Germany);
GESIS – Leibnitz Institute for the Social Sciences, Unter Sachsenhausen 6-8, D-50667 Köln (Germany)

[2] *riechert@forschungsinfo.de*
iFQ – Institute for Research Information and Quality Assurance, Schützenstraße 6a, 10177 Berlin (Germany)

Abstract
Many indicators exist that measure different aspects of scientific productivity, impact, and collaboration. Longitudinal analyses are commonly used to identify developments and changes. However, indicators to quantify dynamics are largely missing and scholarly articles documenting the use of a dynamics indicator are rare. This paper aims at contributing to their development and application. Using *Scopus*, time series of four output indicators in 13 years are observed for 27 disciplines. One qualitative way and two quantitative ways to study dynamics are discussed. The qualitative way is to visualize the data. The first quantitative method to measure growth, the Sharpe Ratio, is imported from portfolio management. The second method is the application of scaling analysis, a way to describe how two properties of a system, like authors and publications, relate to each other as the system undergoes size changes in time. We show that the database is a source of artificial growth, confirming earlier results. Visualizations are an important step to get to know the data, identify potential problems, and generally help interpret quantitative results. The two dynamics indicators reveal different perspectives of growth, but results are correlated with a Pearson coefficient of at least 0.67.

Conference Topic
Scientometrics Indicators - Criticism and new developments (Topic 1), Old and New Data Sources for Scientometric Studies: Coverage, Accuracy and Reliability (Topic 2), and Science Policy and Research Evaluation: Quantitative and Qualitative Approaches (Topic 3).

Introduction
At the same time that the science system experiences growth, globalization, and an increase of interdisciplinary, team, and project work, scientometrics are becoming effectively available. To keep track of developments and analyze its actions, science policy is increasingly interested in scientometric analyses. Many

indicators have been developed and continue to be developed that quantify different aspects of scientific productivity, impact, and collaboration. Longitudinal analyses are commonly used to identify developments and changes. However, indicators to measure dynamics are largely missing. Scholarly articles documenting the use of a dynamics indicator are rare (Grupp *et al.*, 2009). This paper aims at contributing to their development and application.

Using *Scopus*, time series of four output indicators in 13 years are observed for 27 disciplines. One qualitative way and two quantitative ways to study dynamics are discussed. The qualitative way is to visualize the data. Normalizations can be applied to enable comparisons. The first method to quantify growth, the Sharpe Ratio, is imported from portfolio management (Sharpe, 1994). There, it is important to monitor stocks in a portfolio against changes in the whole stock market—a task not so different from science evaluation where database growth must be taken into account. The second method is the application of scaling analysis (Katz, 2000; Lane *et al.*, 2009). The goal is to describe how two properties of a system, like authors and publications, relate to each other as the system undergoes size changes in time. We proceed like Bettencourt *et al.* who applied scaling analysis to measure the dynamics of publications per author (2008).

We start by describing the data and indicators in use. Visualizations are presented as an important step to get to know the data, identify potential problems, and generally help interpret quantitative results. Characteristics, applications, and limitations of the quantitative methods are discussed.

Data and Indicators

We are studying 27 scientific disciplines using the *Scopus* custom database of the German Competence Centre for Bibliometrics. Disciplines are delineated through the *All Science Journal Classification* (ASJC) provided by *Scopus*. The analysis is restricted to the document type "article" and the source type "journal." Publication years 1996 to 2008 are subject to analysis, resulting in 13 years. To avoid the problem of author deflation through homonyms (Strotmann and Zhao, 2012), the author identifier delivered by *Scopus* was used. While this identifier is not sufficiently accurate if authors are the objects of study we deem it accurate enough for macro studies such as ours. We are using disciplines as objects but the instruments to be discussed are applicable to other objects such as fields, countries, or organizations as well.

Indicators are the

- number of publications (P),
- cumulative number of authors (A_{cum}),
- authors per publication (APP), and
- publications per author (PPA).

P and A_{cum} are two measures for the size of a discipline. The assumption behind counting authors cumulatively is that scientists stay in the discipline once they have entered it. APP are calculated on the basis of items, *i.e.*, the number of

distinct author identifiers A per paper P are averaged. PPA is the indicator for productivity. It is the quotient of distinct author identifiers and distinct publications.[151]

Visualizing Growth

The first possibility to study growth is to visualize the data. Figure 1 depicts P and A_{cum} growth curves of seven selected disciplines and total database content. On the level of disciplines, smooth exponential or sigmoid growth is expected (Bettencourt *et al.*, 2008). Instead, in terms of P, we are seeing jumps and dents. The Arts and Humanities jump from 4,000-5,000 publications in 1996-2001 to 7,000-8,000 in 2002-2006. This is because in 2009 the database producer included many new journals that went back as far as 2002.[152] Neuroscience is the only discipline with a publication increase in every year. Almost all disciplines show a decrease of P around the years 2002-2003. The dent is strongly pronounced in the Social Sciences and is clearly visible for the total. This can be explained historically, and reflects the different phases in the creation of *Scopus*.[153]

In general, this fact will need to be dealt with if *Scopus* and the selection of publications through the ASJC are used to measure dynamics. For the purpose of this paper, this complication is instructive. At this point it is clear that database content is a biased estimator of scientific growth.

For the purpose of comparisons it is desirable to show multiple curves in one figure. In practice it is hardly possible to combine more than ten curves without losing comprehensiveness. In addition, it is problematic to combine curves which reside at different scales because the inclusion of curves at a large scale tends to disguise details of curves at small scale. Logarithmic ordinates enable such a combination but also compress the curves and hide details.

Figure 2 (top row) combines growth curves for seven disciplines. Details are already much less visible than in the individual curves of Figure 1. Normalizing growth reintroduces detail (bottom row). To do so, all year values are divided by the initial (1996) value. It is now more easily visible that, in terms of the number of publications, Computer Science, Energy, also the Social Sciences, and Engineering grow stronger in the more recent years than the total which is shown as a solid black line.

To look at productivity (PPA) we must first look at collaboration (APP). Teams are increasingly important in the production of knowledge, most strongly in Science & Engineering, but also in the Social Sciences and the Arts & Humanities (Wuchty *et al.*, 2007). In principle, APP and PPA can grow simultaneously. Consider a field with one publication (having authors X and Y) in one year and

[151] The quotient of maximum author position and distinct publications gives similar results.

[152] Private message from the database producer Elsevier, 14 January 2013.

[153] Document and source types grow differently in Scopus. Elsevier recommends including reviews and conference proceedings articles to arrive at persistent positive growth at the discipline level. Private message from Elsevier, 17 January 2013.

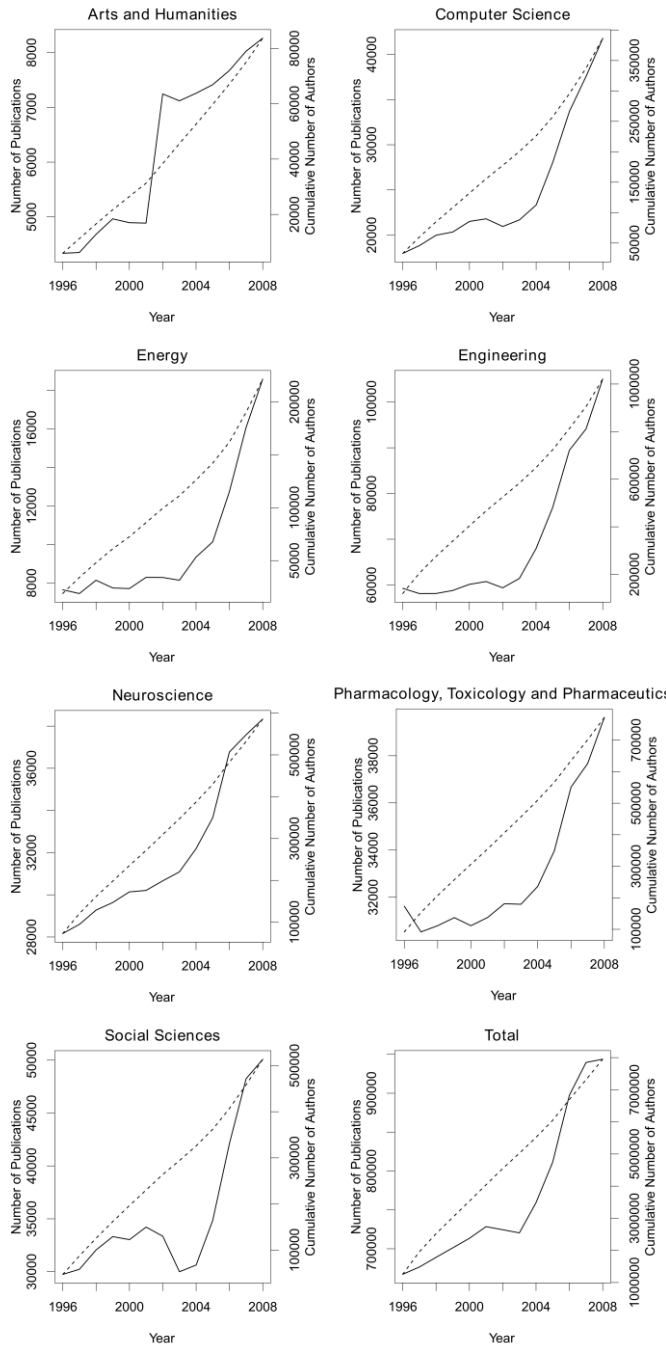


Figure 1. Growth of the number of publications (continuous lines) and the cumulative number of authors (dotted lines) for seven disciplines and total database content in separate figures.

two publications (the first having authors X and Y, the second having authors X, Y, and Z) in the following year. *APP* increases from 2 to 2.5, *PPA* increases from 1/2 to 2/3. Instead, Figure 2 shows that *APP* and *PPA* are inversely proportional when all 13 years are looked at. The increasing dominance of teams prevents a growth of productivity.

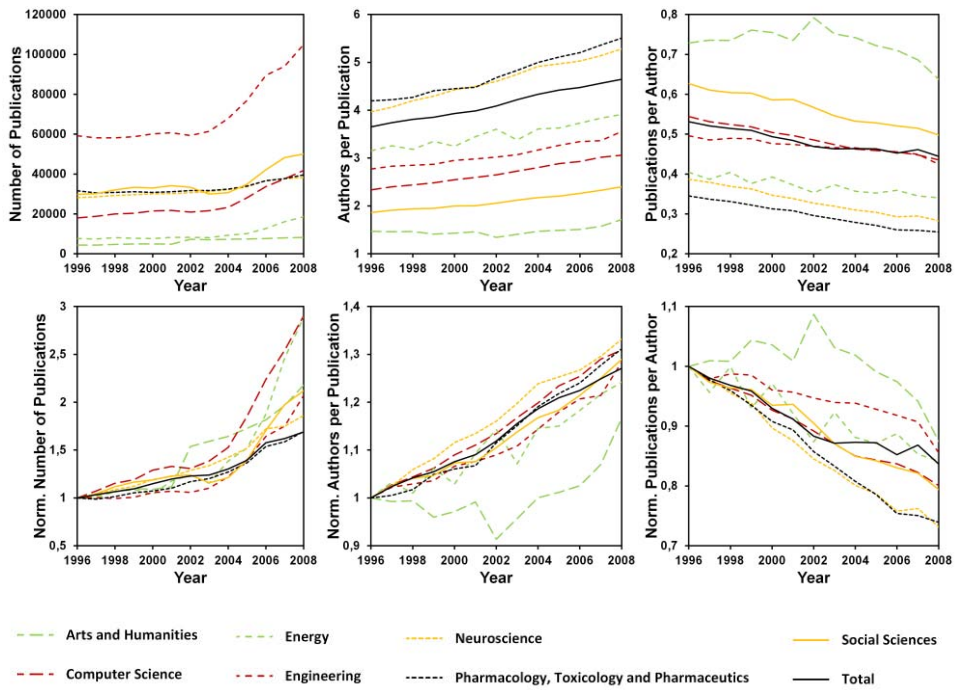


Figure 2. Growth and normalized growth curves of the number of publications, number of authors per publication, and publications per author for seven disciplines and total database content.

Table 1. Sharpe Ratios of the four indicators and scaling exponents of productivity for the whole timespan and the last five years (ranks in brackets).

Discipline	*S(P)*		*S(A_{cum})*		*S(APP)*		*S(PPA)*		*β*	
	13yrs	*5yrs*	*13yrs*	*5yrs*	*13yrs*	*5yrs*	*13yrs*	*5yrs*	*13yrs*	*5yrs*
General	-0.10 (20)	-0.48 (22)	0.84 (18)	2.58 (19)	0.25 (13)	0.20 (20)	-0.37 (14)	-0.35 (10)	0.43 (27)	0.52 (26)
Agricultural and Biological Sciences	0.33 (11)	1.58 (9)	0.96 (12)	4.08 (9)	1.51 (1)	4.16 (1)	-0.76 (22)	-0.73 (15)	0.61 (18)	0.78 (17)
Arts and Humanities	0.22 (15)	-0.62 (24)	1.03 (11)	3.23 (12)	-0.16 (22)	0.71 (11)	0.10 (3)	-1.26 (20)	0.91 (1)	0.48 (27)
Biochemistry, Genetics and Molecular	-0.20 (22)	-0.40 (21)	0.92 (14)	9.46 (3)	0.69 (7)	0.43 (16)	-1.63 (27)	-1.53 (23)	0.55 (22)	0.69 (19)

Biology										
Business, Management and Accounting	0.36 (9)	1.41 (12)	0.88 (16)	11.48 (2)	-0.28 (24)	-2.18 (27)	0.06 (4)	0.95 (1)	0.87 (2)	1.09 (1)
Chemical Engineering	0.64 (2)	0.98 (14)	1.28 (1)	2.58 (18)	0.42 (11)	1.46 (4)	-0.40 (17)	-0.90 (17)	0.77 (8)	0.81 (14)
Chemistry	0.21 (16)	0.31 (18)	1.24 (3)	1.34 (25)	-0.16 (21)	0.96 (9)	-0.07 (9)	-0.44 (12)	0.71 (14)	0.79 (16)
Computer Science	0.96 (1)	4.25 (2)	1.04 (10)	7.42 (5)	0.54 (10)	0.97 (8)	-0.26 (11)	-0.22 (7)	0.81 (7)	0.91 (6)
Decision Sciences	0.34 (10)	2.94 (4)	0.75 (22)	1.83 (22)	-0.46 (27)	-0.05 (23)	0.02 (6)	0.02 (4)	0.82 (6)	0.95 (3)
Dentistry	0.32 (12)	0.47 (16)	1.27 (2)	5.12 (7)	0.72 (6)	0.59 (13)	-0.76 (23)	-1.06 (19)	0.57 (21)	0.76 (18)
Earth and Planetary Sciences	-0.15 (21)	-0.76 (27)	0.38 (26)	0.41 (27)	1.14 (2)	1.13 (5)	-0.38 (15)	-0.51 (13)	0.63 (17)	0.69 (21)
Economics, Econometrics and Finance	0.45 (7)	1.89 (7)	0.75 (21)	2.36 (20)	-0.40 (25)	0.32 (18)	-0.11 (10)	-0.25 (8)	0.76 (9)	0.89 (9)
Energy	0.62 (3)	1.84 (8)	1.08 (6)	2.31 (21)	-0.03 (19)	0.27 (19)	0.03 (5)	-0.02 (5)	0.87 (3)	0.94 (4)
Engineering	0.50 (4)	1.57 (10)	0.82 (19)	2.59 (17)	0.06 (17)	0.64 (12)	0.12 (2)	-0.78 (16)	0.84 (5)	0.84 (12)
Environmental Science	0.10 (17)	1.46 (11)	0.70 (24)	7.46 (4)	0.88 (4)	3.84 (2)	-0.62 (19)	-1.27 (21)	0.72 (12)	0.84 (13)
Health Professions	-0.27 (25)	-0.65 (25)	0.91 (15)	2.93 (14)	-0.41 (26)	-1.49 (26)	-0.38 (16)	-1.60 (24)	0.57 (20)	0.80 (15)
Immunology and Microbiology	-0.30 (26)	-0.53 (23)	0.94 (13)	3.93 (10)	0.59 (9)	0.54 (14)	-1.44 (26)	-1.82 (26)	0.48 (25)	0.66 (25)
Materials Science	0.46 (6)	1.15 (13)	0.70 (23)	1.53 (24)	-0.06 (20)	0.99 (7)	0.21 (1)	-0.15 (6)	0.85 (4)	0.90 (7)
Mathematics	0.42 (8)	2.73 (6)	0.69 (25)	1.62 (23)	0.24 (14)	0.32 (17)	-0.56 (18)	-1.30 (22)	0.61 (19)	0.68 (23)
Medicine	0.30 (14)	3.32 (3)	1.07 (7)	3.08 (13)	0.18 (15)	-0.15 (24)	-0.36 (13)	0.18 (3)	0.66 (16)	0.92 (5)
Neuroscience	-0.22 (24)	-0.71 (26)	1.07 (8)	6.05 (6)	0.60 (8)	0.13 (22)	-1.36 (25)	-2.56 (27)	0.51 (23)	0.69 (20)
Nursing	0.49 (5)	10.61 (1)	1.08 (5)	2.73 (16)	0.83 (5)	0.52 (15)	-0.73 (21)	-0.91 (18)	0.68 (15)	0.89 (10)
Pharmacology, Toxicology and Pharmaceutics	-0.42 (27)	-0.17 (20)	1.06 (9)	12.98 (1)	0.25 (12)	2.50 (3)	-0.72 (20)	-0.61 (14)	0.47 (26)	0.69 (22)
Physics and Astronomy	-0.07 (19)	0.46 (17)	-0.71 (27)	0.49 (26)	-0.24 (23)	0.19 (21)	-0.03 (7)	-0.27 (9)	0.75 (10)	0.85 (11)
Psychology	0.08 (18)	0.84 (15)	0.86 (17)	3.37 (11)	0.06 (18)	-0.40 (25)	-0.03 (8)	0.38 (2)	0.71 (13)	1.03 (2)
Social Sciences	0.32 (13)	2.77 (5)	0.78 (20)	2.92 (15)	0.17 (16)	0.95 (10)	-0.29 (12)	-0.43 (11)	0.74 (11)	0.90 (8)
Veterinary	-0.20 (23)	-0.17 (19)	1.11 (4)	4.73 (8)	0.94 (3)	1.07 (6)	-0.83 (24)	-1.63 (25)	0.48 (24)	0.67 (24)
Total	—	—	—	—	—	—	—	—	0.71	0.88

While figures of normalized variables obscure dimensions, they are capable of delivering basic growth messages in a qualitative way. When quantifying growth two things should be kept in mind. First, because database content is itself

dynamic indicators should correct for that. Second, due to changes in growth it may be necessary to compute growth indicators for different time regimes. In our case, because of the dynamics of the *Scopus* database itself, we will study indicators for all 13 years and just the last five years where dynamics are more stable.

The Sharpe Ratio

The Sharpe Ratio is a metric from financial portfolio management, a measure of average annual growth (Sharpe, 1994). It has been applied in science policy contexts to identify key institutions (Schmoch *et al.*, 2006) and dynamic research fields (Grupp *et al.*, 2009). Although in these and other studies the metric is applied to raw publication counts, in principle it can be applied to describe the dynamics of all indicators. The (historic) Sharpe Ratio of a quantity N_i in discipline i is defined as

$$S(N_i) = \frac{\overline{D_i(t)}}{\sigma_{D_i(t)}} \tag{1}$$

with

$$D_i(t) = \frac{N_i(t+1) - N_i(t)}{N_i(t)} - \frac{N_{total}(t+1) - N_{total}(t)}{N_{total}(t)} \tag{2}$$

where $\overline{D_i(t)}$ is the average of $D_i(t)$ and $\sigma_{D_i(t)}$ is its standard deviation. In words, first, the annual growth rate of the desired quantity in discipline i is calculated for each year. Second, to normalize for database growth the annual growth rate of the total is subtracted, giving the normalized annual growth rate. Third, the normalized annual growth rate is averaged over all years, giving the average normalized growth rate. Fourth, the average normalized growth rate is divided by the standard deviation of the normalized annual growth rates. Since the standard deviation is small when the normalized annual growth rates do not vary much, steady growth is rewarded and erratic growth is punished.

Table 1 gives the Sharpe Ratios for all 13 years and just the last five years. S is not intuitively interpretable. A negative value does not imply negative growth, it means that a discipline grows less than the total. The indicator confirms the impressions from Figure 2. Computer Science, Energy, and Engineering have the strongest publication growth. Pharmacology, Toxicology and Pharmaceutics has a negative Sharpe Ratio and occupies the last rank position. Because the Arts and Humanities hardly grow except for the artificial jump, they take almost last rank when just the last five years are considered. The top rank for Nursing is because the number of publications more than doubled in the last five years which again is an artifact of database production.

Regarding productivity, the three top curves of Figure 2, Engineering (0.12), Arts and Humanities (0.10), and Energy (0.03) are among the disciplines with the highest Sharpe Ratios. Pharmacology, Toxicology and Pharmaceutics (-0.72) as well as Neuroscience (-1.36) grow much less than the total, as can be seen in the figure. When just the last five years are considered, the Arts and Humanities fall 17 rank positions, reflecting that the artifact does not influence the score anymore. It may be surprising that only six scientific disciplines exhibit a growth of productivity. Again, we look at its relation to cooperation, now in a quantitative way. Engineering is the only discipline with both cooperation and productivity growth. The average growth of cooperation in all 27 disciplines over 13 years is 0.29 while average productivity growth is -0.43. The Pearson correlation coefficient of both quantities is -0.64. This confirms the impression from Figure 2 that, overall, the increasing dominance of teams indeed prevents a growth of productivity. And teams do become more important. Cooperation growth is even bigger (0.69) when just the last five years are studied. Recent 5 year average cooperation growth is 137% bigger than overall 13 year growth. Productivity, on the other hand, does not decrease as drastically. Recent 5 year average productivity growth (-0.71) is just 65% bigger than 13 year growth. In the 5 year window, correlation of both quantities gets lost (the Pearson coefficient is -0.24). This may indicate that science is entering a phase where productivity grows despite growing cooperation, but confidence in this statement is muddled due to the database artifact we have found.

Scaling Analysis

Another way to quantify dynamics is scaling analysis, a method to study the behaviour of a complex system across spatial or temporal scales (Lane *et al.*, 2009). Scaling analysis belongs to modeling but can also be used for evaluation purposes. A system is said to scale when two properties N and Y are related through a power law $Y \propto N^{\alpha}$. The scaling exponent α is then a system-specific descriptor of the average relative change in Y with N. If $\alpha > 1$, Y/N increases with N (increasing returns). On the contrary, if $\alpha < 1$, Y/N decreases in an economy of scale. For $\alpha = 1$, Y/N is constant.

Originating in biology, astrophysics, and urban studies, scaling analysis was later applied to science and innovation systems and it was demonstrated that different scientific disciplines all share the property of scale invariance, making the method applicable to systems with different publication behaviors (Katz, 1999, 2000). Bettencourt *et al.* (2008) have applied scaling analysis to model science system dynamics. If N is the number of authors A at time t and Y is the number of publications P at that time, $\beta > 1$ ($\beta < 1$) characterizes a research field with increasing (decreasing) individual productivity (PPA):

$$P_i(t) \propto A_i(t)^{\beta} \tag{3}$$

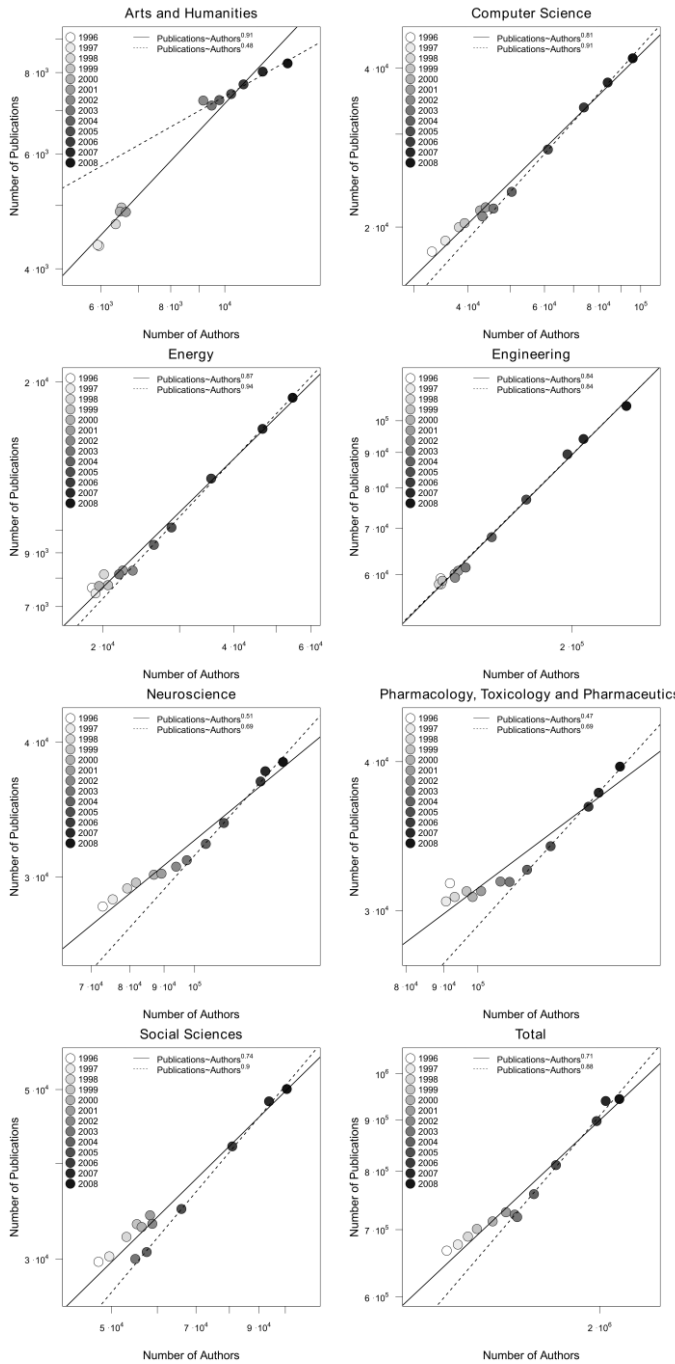


Figure 3. Scaling analysis (productivity) of seven disciplines and total database content (the continuous line is a fit to all data points, the dotted line to just the last 5 years).

We have used standardized major axis analysis, implemented in R, to estimate the scaling exponent β because the method is scale invariant and we are not interested in inference (Warton *et al.*, 2006, 2012).
Results are given in Figure 3 and Table 1. Each data point in a figure corresponds to a year. Grayscales are used to mark years. If both A and P increase monotonously from year to year data points monotonously move from the bottom left to the upper right corner as time passes. This is only the case for Neuroscience. But even there different growth regimes are visible as non-parallel lines. Continuous lines are fits to all 13 years, dotted lines just to the last five years. The artifact caused by database growth is visible in most disciplines, also for the total database content. In the Social Sciences the second regime starts at a scale way below of what had been reached in the first regime. In the Arts and Humanities it is now clearly visible that in the last five years growth is much smaller than for the system as a whole.
Exponents for all disciplines are smaller than 1, saying that productivity decreases as size increases. Top disciplines are those with largest exponents. Again, the Arts and Humanities (0.90), Energy (0.87), and Engineering (0.84) have top scores while Pharmacology, Toxicology and Pharmaceutics (0.47) and Neuroscience (0.51) are at the lower end of productivity. When just considering the last five years, the Arts and Humanities are punished more than if the Sharpe Ratio is used. They drop from first to last rank position.
Even though β and S are different perspectives on growth, the two indicators are quite strongly correlated. Pearson correlation coefficients are 0.77 for the whole timespan and 0.67 for the last five years. This is because β is mathematically related to changes in Y/N, the annual growth rate, that S is based on.

Discussion and Conclusion

Starting from a need for metrics of dynamics in the science system, visual growth curves, the Sharpe Ratio, and scaling exponents were discussed. Curves can unveil essential meanings in a qualitative way. Comparison of objects of study is made easier by normalizing curves. In our case, the visualization revealed different growth regimes where just one regime was expected, showing that the database is a source of artificial growth, confirming earlier results (Larsen and Ins, 2010; Michels and Schmoch, 2012). Knowing the data is thus imperative and actually looking at it should be a first step before growth is quantified.
Two metrics, the Sharpe Ratio S and scaling exponent β, were discussed and applied to whole and partial timespans to cope with the presence of different growth regimes. The Sharpe Ratio is based on the average annual growth rate and aims to remove artifacts by subtracting overall database growth. On the level of this study, this subtraction may be criticized because overall database content is dominated by the hard sciences but is also used to normalize the soft sciences. But once objects in a coherent field are studied, like countries in a discipline, such a normalization immediately makes sense. In addition, the Sharpe Ratio addresses

the stability of growth by dividing by the standard deviation of the normalized annual growth rates.

In scaling analysis, visualizations also proved to be important as they made different regimes very transparent and stressed the need to fit different functions to them (Bettencourt *et al.*, 2008). A spontaneous reaction is to criticize fits to just five data points. Of course, the fewer data points there are, the more each one influences the result. But this is also true for the average growth rate which is frequently used also for small numbers. Scaling exponents are not normalized for database dynamics, but a way to normalize would be to divide by the exponent of the total.

The Sharpe Ratio can be applied to all possible time series, not just to publication counts, as common in the literature. If one does not feel comfortable with dividing by the standard deviation, this can easily be left out. Scaling analysis is only applicable to bivariate data. Besides productivity another natural application is to quantify impact dynamics (Katz, 2000).

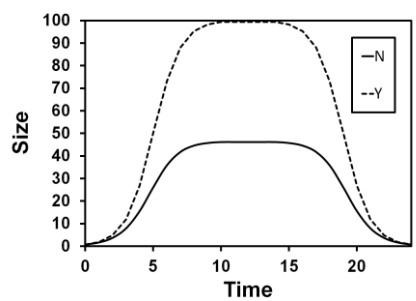


Figure 4. Growth dynamics of an hypothetical system with $Y \propto N^{1.2}$.

Scale only corresponds to time if annual growth rates are constantly positive. In other words, time does not move backwards when variables decrease. Consider the system depicted in Figure 4. It has the typical dynamics of an emerging, then persisting, and finally dying field with $Y \propto N^{1.2}$. If all years are subject to scaling analysis, the exponent will still be 1.2. If scaling analysis is done like it is done here, the only sign that the field is actually shrinking is that the data points become darker as they move from the upper right corner to the lower left. Most obvious applications of scaling analysis may be the identification and characterization of emerging science (Guo *et al.*, 2011).

To conclude, the Sharpe Ratio and scaling exponents are different perspectives on growth with different areas of applicability. They should be used in combination with growth visualizations which help get a feeling for the data and which can reveal intricacies of the system under study. A result of this technical discussion that requires further scrutiny is that the increasing dominance of teams in the production of knowledge actually prevents a growth of productivity, less so in the last five than in the last 13 years.

Acknowledgments

Thanks to Almuth Lietz (iFQ) for help with the figures.

References

Bettencourt, L.M.A., Kaiser, D.I., Kaur, J., Castillo-Chávez, C. & Wojick, D.E. (2008). Population modeling of the emergence and development of scientific fields. *Scientometrics*, 75, 495-518.

Grupp, H., Hinze, S., Breitschopf, B. (2009). Defining regional research priorities: a new approach. *Science and Public Policy*, 36, 549-559.

Guo, H., Weingart, S. & Börner, K. (2011). Mixed-indicators model for identifying emerging research areas. *Scientometrics*, 89, 421-435.

Katz, J.S. (1999). The self-similar science system. *Research Policy*, 28, 501-517.

Katz, J.S. (2000). Scale-independent indicators and research evaluation. *Science and Public Policy*, 27, 23-36.

Lane, D., Pumain, D. & Leeuw, S. van der (Eds.) (2009). *Complexity Perspectives in Innovation and Social Change*. Dordrecht: Springer.

Larsen, P.O. & Ins, M. von (2010). The rate of growth in scientific publication and the decline in coverage provided by Science Citation Index. *Scientometrics*, 84, 575-603.

Michels, C. & Schmoch, U. (2012). The growth of science and database coverage. *Scientometrics*, 93, 831-846.

Schmoch, U., Wang, J. & Stoica, R. (2006). *Research and Development in the Turkish Landscape of Science: Identification of Key Institutions*. Report to the Federal Ministry of Education and Research (BMBF), Fraunhofer Institute for Systems and Innovation Research.

Sharpe, W.F. (1994). The Sharpe Ratio. *The Journal of Portfolio Management*, 21, 49-58.

Strotmann, A. & Zhao, D. (2012). Author name disambiguation: What difference does it make in author-based citation analysis? *Journal of the American Society for Information Science and Technology*, 63, 1820-1833.

Warton, D.I., Duursma, R.A., Falster, D.S. & Taskinen, S. (2012). smatr 3- an R package for estimation and inference about allometric lines. *Methods in Ecology and Evolution*, 3, 257-259.

Warton, D.I., Wright, I.J., Falster, D.S. & Westoby, M. (2006). Bivariate line-fitting methods for allometry. *Biological Reviews*, 81, 259-291.

Wuchty, S., Jones, B.F. & Uzzi, B. (2007). The increasing dominance of teams in production of knowledge. *Science*, 316, 1036-1039.

SCIENTIFIC POLICY IN BRAZIL: EXPLORATORY ANALYSIS OF ASSESSMENT CRITERIA (RIP)

Elaine Cristina Pinto de Miranda[1], Rogério Mugnaini[1,2],

[1] *elainecpm@hotmail.com*
Universidade de São Paulo, Escola de Comunicações e Artes, Depto. de Ciência da Informação, Av. Prof. Lúcio Martins Rodrigues, 443, CEP 05508-020, São Paulo, SP, Brazil

[2] *mugnaini@usp.br*
Universidade de São Paulo, Escola de Artes, Ciências e Humanidades, Av. Arlindo Bettio, 1000, CEP 03828-000,São Paulo, SP, Brazil

Abstract
Brazilian scientific evaluation process involves nowadays 3,000 postgraduate programs and almost 1,000 of *ad hoc* consultants from educational institutions of all regions of the country. Each triennial cycle, produces a lot of information, part of it qualitative, derived from the decisions of the committees of the 46 assessment areas. Our study proposes a documentary analysis of this documentation, in order to express contextual aspects of scientific communication process in each area. We aimed to compare the different scientific fields, due to the importance given to publication in scientific journals as well as its contrast to books and national journals importance. The results here presented show groups of Assessment Areas at different stages of the scientific communication process. We found that there are areas in which publications occurs primarily in indexed international journals, otherwise there are areas proposing specific criteria to evaluate the quality of national journals. There are also areas that are in the process of establishing their journals, and others are being forced to change the practice of publishing books and start publishing on journals.

Conference Topic
Science Policy and Research Evaluation: Quantitative and Qualitative Approaches (Topic 3)

Introduction
Since 1976, the Coordination of Higher Education Personnel Improvement (CAPES) is developing a national wide evaluation process of Brazilian postgraduate programs, which has been significantly improved between 1996 and 1997, in order to establish a triennial evaluation cycle since 1998. This is a huge effort considering that Brazil presents nowadays around 3,000 postgraduate programs. The evaluation process involves almost 1,000 of *ad hoc* consultants from educational institutions of all regions of the country, composing committees in 46 assessment areas of the different scientific fields[154]. Between the different

[154] http://www.capes.gov.br/avaliacao/tabela-de-areas-de-conhecimento

aspects evaluated, scientific production is the one that most influence and determine the level of the postgraduate program (Souza & Paula, 2002).

Follow up evaluations are done annually, and the final one, at the end of each triennial cycle, producing a lot of information published in the Coordination's website: documents, tables and even raw bibliographic data. The information can be found in different levels of aggregation (researcher, postgraduate program, institution, assessment area and broad scientific area).

Significant part of the information gathered by CAPES comes directly from Lattes Platform[155] – the curriculum database maintained by the National Council for Scientific and Technological Development (CNPq). The reliability of this process is based on policies that guarantee that the researchers will maintain their curriculum up-to-date periodically. It will refer about all information related to the scientific production, research projects, academic career, and any information inserted in the curriculum. The objective character of this type of information has served a diversity of scientometric analysis that has been widely published in important specialized information sources (Leite et al., 2011).

The other part consists mainly of qualitative information, which composes the documentation derived from the decisions of the committees (CAPES, 2011), in order to define specific criteria to each assessment area. These committees are composed by the Area Representative at CAPES and *ad hoc* consultants, who are responsible for setting criteria for qualifying vehicles and their classification for subsequent use as an input to postgraduate program evaluation (Souza & Paula, 2002). This rich information reflects the consensus among the researchers, bringing in itself, the capability of express contextual aspects of scientific communication process in each area.

Hicks (2004) argues that, although journal articles have its importance, in Social Sciences and Humanities, books publishing predominates, and due to its characteristics an ideal evaluation should consider what she calls the four literatures: journals, books, national and non-scholarly literature – being "national literature" the one developed in local context and "non-scholarly literature" the knowledge reaching out to application.

These concerns are contemplated in CAPES documentation, which requires each area decide about setting specific criteria based in: (1) a book classification form, and; (2) a national citation index to infer quality of national journals.

This study aims to compare the difference between the scientific communication process in the 46 areas of CAPES national assessment, based in the documentation derived from the decisions of the committees of each assessment area. More specifically it analyses the importance given to publication in scientific journals as well as its contrast to books and national journals importance. The next stage of this ongoing project aims to evaluate the whole Brazilian Scientific Community, including the scientific indicators derived from the Lattes Platform.

[155] http://www.capes.gov.br/avaliacao/coleta-de-dados

Methodology

An exploratory research was carried out, applying documentary analysis to the documents proposed by the committees of the 46 assessment areas, classified into nine broad subject areas, that are described on Figure 1. These documents are part of the Qualis, a set of procedures used by CAPES to realize a more systematic treatment and quality of scientific intellectual output of postgraduate programs aiming to improve the indicators that support the evaluation of these programs (Souza & Paula, 2002).

Agricultural Sciences	
AGRICULTURAL SCIENCES	AGRIC
ANIMAL SCIENCE / FISHING RESOURCES	AN-SCI, FISH
FOOD SCIENCE AND TECHNOLOGY	FOOD S&T
VETERINARY MEDICINE	VET-MED
Applied Social Sciences	
APPLIED SOCIAL SCIENCES	APP-SOC-SCI
ARCHITECTURE, URBAN PLANNING AND DESIGN	ARCHT, URB, DESN
BUSINESS, ACCOUNTING (SCIENCES) AND TOURISM	BUS, ACC, TOUR
ECONOMICS	ECON
LAW	LAW
SOCIAL SERVICES / DOMESTIC ECONOMY	SOC-SERV, DOM-ECON
URBAN AND REGIONAL PLANNING / DEMOGRAPHICS	URB-REG-PLAN, DEMOG
Biological Sciences	
BIOLOGICAL SCIENCES I	BIO 1
BIOLOGICAL SCIENCES II	BIO2
BIOLOGICAL SCIENCES III	BIO 3
ECOLOGY AND ENVIRONMENT	ECOL, ENV

Engineering	
ENGINEERING I	ENG 1
ENGINEERING II	ENG 2
ENGINEERING III	ENG 3
ENGINEERING IV	ENG 4
Exact and Earth Sciences	
ASTRONOMY / PHYSICS	ASTR, PHYS
CHEMISTRY	CHEM
COMPUTER SCIENCE	COMP
GEOSCIENCES	GEOSC
MATH / PROBABILITY AND STATISTICS	MATH, PRO
Health Sciences	
DENTISTRY	DENT
MEDICINE I	MED 1
MEDICINE II	MED 2
MEDICINE III	MED 3
NURSING	NURS
PHARMACY	PHARM
PHYSICAL EDUCATION	PHYS-EDUC
PUBLIC HEALTH	PUB-HEAL

	Engineering		**Human Sciences**	
AGRIC	ENGINEERING I	ENG 1	ANTHROPOLOGY / ARCHAEOLOGY	ANTR, ARCH
AN-SCI, FISH	ENGINEERING II	ENG 2	EDUCATION	EDUC
OOD S&T	ENGINEERING III	ENG 3	GEOGRAPHY	GEOGR
VET-MED	ENGINEERING IV	ENG 4	HISTORY	HIST
	Exact and Earth Sciences		PHILOSOPHY / THEOLOGY	PHIL, THEOL
APP-SOC-SCI	ASTRONOMY / PHYSICS	ASTR, PHYS	POLITICAL SCIENCE AND INTERNATIONAL RELATIONS	POL-SCI, INT-REL
ARCHT, URB, DESN	CHEMISTRY	CHEM	PSYCHOLOGY	PSYCH
BUS, ACC, TOUR	COMPUTER SCIENCE	COMP	SOCIOLOGY	SOCIOL
ECON	GEOSCIENCES	GEOSC	**Linguistics, Letters and Arts**	
LAW	MATH / PROBABILITY AND STATISTICS	MATH, PROB, STAT	ARTS / MUSIC	ART, MUSC
OC-SERV, DOM-ECON	**Health Sciences**		LETTERS / LINGUISTICS	LETT. LING
URB-REG-PLAN, DEMOG	DENTISTRY	DENT	**Multidisciplinar**	
	MEDICINE I	MED 1	BIOTECHNOLOGY	BIOTECH
BIO 1	MEDICINE II	MED 2	INTERDISCIPLINARY	INTERD
BIO2	MEDICINE III	MED 3	MATERIALS	MATER-SCI
BIO 3	NURSING	NURS	SCIENCE AND MATH TEACHING	SCI&MATH -TEACH
ECOL, ENV	PHARMACY	PHARM		
	PHYSICAL EDUCATION	PHYS-EDUC		
	PUBLIC HEALTH	PUB-HEAL		

Figure 1: List of acronyms of Assessment Areas and respective Broad Areas

Qualis consists of the classification of vehicles in which the scholars of postgraduate programs publish their findings. Titles are analyzed and ranked according to criteria established by committees of area. According to Capes (2004), "is at the discretion of each area to decide on the category of vehicle used

by it: there are areas that rank only journals, as there are those who classify other types of vehicles such as: annals, newspapers and magazines". As one might expect, the relative importance of journals, against books and proceedings is a constant discussion of the committees of different assessment areas, that develops its own document based on the main guidelines from CAPES Board of Assessment. These documents are reviewed and approved by the Scientific Technical Council (CTC).

The **CAPES document** is structured into six parts: a) Identification (Assessment Area, area coordinator, area assistant coordinator, modality); b) I. General considerations on the current stage of area; c) II. General considerations on the evaluation form for the respective triennium; d) III. General Considerations for Qualis journals, books classification form and criteria for usage and stratification of these documents in the assessment; IV. General Evaluation Form for the respective triennium; e) V. Considerations and definitions about assigning grades 6 and 7 (maximum) – international insertion.

As this study intent to study the specific criteria to evaluate books and national journals, it considered the **sections III** and **IV** that are described below, from the last completed triennial assessment (2007 - 2009).

Section III: consists in the defined criteria to (1) journals classification (A1, A2, B1, B2, B3, B4, B5 – where A1 and A2 are the core journals of each area, and C means "no value"). At one hand, the majority of the areas consider the journal Impact Factor from the Journal Citation Reports (JCR IF) as the main and unique indicator to define the core journals, while other areas value national journals indexed at SciELO (Scientific Electronic Library Online) database. On the other hand, some areas consider the book in its evaluation and also describe the books classification in a detailed (complex and controversial) way, defining the book classification form. A couple of areas still classify conference proceedings, even calculating specific indicators to do so - that is the case of Computer Science.

Section IV: shows the general *Evaluation Form*. It is composed by five items and sub items that receive a weight for the whole evaluation, for example, in a specific area: 1. Proposal of the postgraduate program (0%); 2. Docent staff (20%); 3. Students Thesis and Dissertations (35%); 4. Intellectual Production (35%); 5. Social Inclusion (10%). Each area has the autonomy to set the weight of the items and sub items, in a weight range established at the main guidelines from CAPES Board of Assessment.

In order to map the diverse attribution of relevance of scientific journals and books between the different assessment areas, the respective weights were analyzed (information gathered from item 4. Intellectual Production of *Evaluation Form*). The documents were obtained from CAPES Board of Assessment that sent the material on December 2012.

Information of interest were: Broad Area (BA), Assessment Area (AA), the definition or not, by the AA of a detailed book classification form and the consideration about the presence of a journal at SciELO database as a quality criteria. As response variables were selected the JCR IF and the weight attributed

to publication on scientific journal (that multiplied by the weight associated to Intellectual Production, gives the importance of publication in journal articles against all the other aspects of the general Evaluation Form).

Findings and discussion

The first variable analyzed concerns to the percentual weight of the publication on journals in relation to all aspects considered in the general Evaluation Form.

Observing Figure 2 it is possible to see that the average weight assigned by the AAs of each BA is more significant (> 22%) for EXA and AGR, followed by SOC, ENG, MULT and HLTH (around 18%), and HUM e BIO (around 17.5%) and LLA (16%). The BAs showing more AAs (8) are HLTH and HUM.

The data shows (figure not presented) that most AAs (58.6%) use books classification criteria and, contrarily, the majority (63,6%) do not consider being indexed in SciELO as a criteria for the journal classification on the highest stratum. But is interesting to note that SciELO is more used in areas that classify books (41.4%) than in areas that don't classify (29.4%).

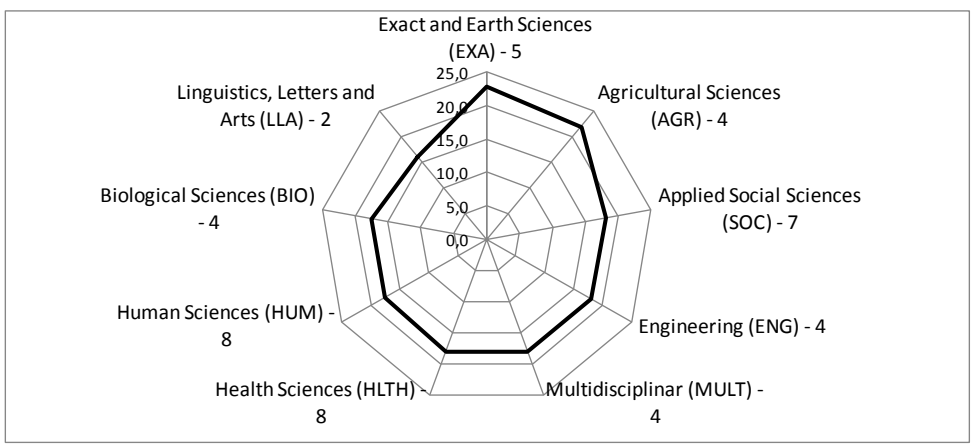


Figure 2. Broad Areas (and respective number of Assessment Areas) distributed by descendent average weight attributed to journal in the postgraduate program assessment.

Another aspect to be mentioned is that the 17 AAs that do not classify books are mostly the so called hard sciences, assigning an average weight to journals in 20.3%. However, when considering the 29 AAs that classify books, the 5 Social Sciences areas attribute the highest average weight to journals (18.6%), followed by 9 hard sciences areas (18.0%) and finally by 3 Arts and Humanities areas (15.1%).

Table 1 show areas that don't use books classification criteria or SciELO indexing to classify journals on the highest stratum (Group 2) are those with highest weight for journal publications. This group concentrates 60% of EXA AAs and 50% of

HLTH. It is the second group that requires higher average value of JCR IF to be at the highest stratum, presenting the largest JCR IF on BIO (*BIO 3*, with 4.9) and AGR (*VET-MED*, with 2.6).

Areas that do not value books but consider indexing in SciELO as criteria for evaluating journals (Group 1) are those that require higher JCR IF as evaluation criteria of journals on the top stratum. We highlight the *ASTR, PHYS* require JCR IF 6.0 (the highest of EXA) and *ENG 4* (IF 1.0) with the higher requirement across the ENG fields. The group concentrates 40% of EXA AAs and 50% of ENG. It should be emphasized that the presence ENG and *COMP* on this group is because they are areas that use mostly conference proceedings to disseminate their results, not considering either national journals or books – as consequence they require low JCR IF to classify journals.

Group 3 shows AAs that classify books but do not consider SciELO indexing as evaluation criteria. It is possible to see that it is the most heterogeneous group and contains the majority of the AAs. Displays the *BIOTECH* area which requires JCR IF 5.0 (the highest into MULT).

And Group 4 concentrates the most part of AAs in HUM (75%) and SOC (57%), and classifies books and considers indexing in SciELO as evaluation criteria. These areas do not often use the JCR IF as journals evaluation criteria on the highest stratum, except in *GEOGR* areas (IF of 0.5) and *PUB-HEAL* (IF of 4.0, the highest of HLTH).

The evaluation of scientific production nationwide should consider the existing policy model as well as add the scientific communication specificities expressed in the practices of communities of different knowledge areas. The evaluation criteria should be appropriate to the different areas, and also to the national context, in order to give subsidies to a coherent science policy, in which the Brazilian scientific community has played an important role.

As explains Trigueiro (2001), the scientific community, together with the state, contributed significantly to the establishment and consolidation of scientific and technological national base through scientific societies that now use the policy weapons, establishing: the institutions that support research, postgraduate programs, national plans and the current evaluation system itself. Already Guédon identifies the emergence of a power structure formed at the scientific field, whose components are institutions, associations and journals - that is involved in a dense and complex web of interactions and influences, characterizing the scientific field. "Institutions, associations and journals will also be relevant to any study of power and competition in the social sciences and the humanities, but they will not work in the same way as in science. Together, they form a national system of science" (Guédon, 2010, p. 26).

At this context Qualis evaluation process is being developed, while it is always trying to adjust the criteria to the specificities of each area, but always seeking to establish a high standard of excellence, aiming to lead scientific production to most qualified vehicles.

Table 2. Distribution of Assessment Areas (and respective Broad Area) by journal weight and Impact Factor (JCR) required to get the highest stratum classification, grouped by usage or not of book classification form and/or national journal SciELO indexing

Group 1: No books | Yes SciELO

Broad Area		Assessment Area	Journal weight (%)	IF JCR
Area	%			
BIO	*25%*	BIO2	14	4.7
EXA	*40%*	ASTR, PHYS	17.5	6
		COMP	26	1.4
ENG	*50%*	ENG 4	17.5	1
		ENG 3	17.5	-
		Avarage	**18.5**	**3.3**

Group 2: No books | No SciELO

Broad Area		Assessment Area	Journal weight (%)	IF JCR
Area	%			
BIO	*25%*	BIO 3	20	4.9
HLTH	*50%*	MED 2	20	3.8
		MED 1	20	3.8
		DENT	20	3.1
		MED 3	20	3
EXA	*60%*	CHEM	21	4
		GEOSC	24	2.8
		MATH, PROB, STAT	26	1
AGR	*50%*	VET-MED	22	2.6
		AGRIC	22	2
MULT	*25%*	MATER-SCI	17.5	1
ENG	*25%*	ENG 2	20	1
		Avarage	**21.0**	**2.7**

Group 3: Yes books | No SciELO

Broad Area		Assessment Area	Journal weight (%)	IF JCR
Area	%			
MULT	*50%*	BIOTECH	16	5
		SCI&MATH -TEACH	17.5	-
BIO	*50%*	BIO 1	15.8	4.1
		ECOL, ENV	20	3
AGR	*50%*	FOOD S&T	22	2.6
		AN-SCI, FISH	22	2
HLTH	*38%*	PHARM	16	3
		PHYS-EDUC	16	1.9
		NURS	16	0.8
ENG	*25%*	ENG 1	17.5	0.8
SOC	*43%*	BUS, ACC, TOUR	22.8	0.5
		ECON	22.8	-
		ARCHT, URB, DESN	16	-
HUM	*25%*	PHIL, THEOL	17.5	-
		EDUC	17.5	-
LLA	*100%*	LETT. LING	20	-
		ART, MUSC	12	-
		Avarage	**18.1**	**2.4**

Group 4: Yes books | Yes SciELO

Broad Area		Assessment Area	Journal weight (%)	IF JCR
Area	%			
HLTH	*13%*	PUB-HEAL	16	4
HUM	*75%*	GEOGR	14	0.5
		PSYCH	17.5	-
		POL-SCI, INT-REL	24	-
		SOCIOL	20	-
		ANTR, ARCH	16	-
		HIST	14	-
SOC	*57%*	URB&REG-PLAN, DEMOG	17.5	-
		SOC-SERV, DOM-ECON	16	-
		APP-SOC-SCI	16	-
		LAW	16	-
MULT	*25%*	INTERD	21	-
		Avarage	**17.3**	**2.3**

Currently it is required that the higher stratum of the journals (A1 and A2) must contain no more than 25% of the journals classified by the area and the number of journals comprising the stratum A1 must be smaller than those classified as A2. This restriction was promptly opposed by several areas - taking as an example the Public Health, because two very important national journals that are doomed to second stratum (A2), what is a problem in one area in which large percentage of the production is directed to the national reality and has no significant international impact (ABRASCO, 2008). However the last triennial evaluation document (2007-2009) shows a remarkable advance by considering national journals indexed in SciELO, as we observed on this study, mostly on Group 4.

The outcome is in line with the evaluation studies carried out in some countries where there is also a preoccupation with differences between the areas and the different forms of communication. Among then we highlight Larivière et al (2006) who claim that "while the validity and appropriateness of bibliometric methods are largely accepted in the natural sciences, the situation is more complex in the case of the social sciences and humanities". To the authors, "evaluations based only on measures obtained from journal databases are more likely to be less than adequate for disciplines in which less than 50% of references are made to journal articles than for those in which these references account for more than 50%".

Final remarks

The results here presented show groups of AAs at different stages of the scientific communication process. There are areas in which publications occurs primarily on indexed international journals, otherwise there are areas proposing specific criteria to evaluate the quality of national journals. There are also areas that are in the process of establishing their journals, and others being forced to publish on journals.

Various adjustments can be observed in passing periods, and a longitudinal approach of this analysis will be undertaken soon.

Acknowledgments

We acknowledge FAPESP, that supports the *Young Investigators Awards* project, intitled *Scientific assessment in Brazil: study of scientific communication in scientific areas* - Grant number 2012/00255-6[156].

References

ABRASCO (2008). Nota do fórum de coordenadores de Programas de Pós-Graduação em saúde coletiva sobre o novo Qualis Periódicos. *Saúde e Sociedade*, 17 (4).

CAPES - Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (2004). QUALIS: Concepção e diretrizes básicas. *Revista Brasileira de Pós-Graduação*, 1 (1), 149-151. Retrieved January 14, 2013 from: http://www2.capes.gov.br/rbpg/images/stories/downloads/RBPG/Vol.1_1_jul2004_/Qualis_ConcepcaoDiretrizes.pdf.

CAPES - Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (2011). *Critérios de avaliação*. Retrieved December 7, 2011 from: http://www.capes.gov.br/avaliacao/criterios-de-avaliacao.

Guédon, J. C. (2010). Acesso Aberto e a divisão entre ciência predominante e ciência periférica. In: Ferreira, S.M.S.P. & Targino, M.G. (Orgs). *Acessibilidade e visibilidade de revistas científicas eletrônicas* (pp. 21-77). São Paulo: Senac/CENGAGE Learning.

[156] http://www.bv.fapesp.br/en/auxilios/48066/scientific-assessment-brazil-study-scientific/

Hicks, D. (2004). *The four literatures of Social Science*. Retrieved January 12, 2013 from: http://works.bepress.com/diana_hicks/16.

Larivière, V. et al. (2006). The Place of Serials in Referencing Practices: Comparing Natural Sciences and Engineering with Social Sciences and Humanities. Journal of the American Society for Information Science and Technology, 57 (8), 997–1004.

Leite, P.; Mugnaini, R.; Leta, J. (2011). A new indicator for international visibility: exploring Brazilian scientific community. *Scientometrics*. 88: 311-319.

Souza E. P.; Paula M. C. S. (2002). QUALIS: a base de qualificação dos periódicos científicos utilizada na avaliação CAPES. *INFOCAPES – Boletim Informativo da CAPES*, 10 (2), 7-25. Retrieved August 1, 2011 from: http://www.capes.gov.br/images/stories/download/bolsas/Infocapes10_2_2002.pdf.

Trigueiro, M. G. S. (2001). A comunidade científica, o Estado e as universidades, no atual estágio de desenvolvimento científico tecnológico. *Sociologias*, 6, 30-50.

‘SEED+EXPAND’: A VALIDATED METHODOLOGY FOR CREATING HIGH QUALITY PUBLICATION OEUVRES OF INDIVIDUAL RESEARCHERS

Linda Reijnhoudt[1], Rodrigo Costas[2], Ed Noyons[2], Katy Börner[1,3], Andrea Scharnhorst[1]

[1] *linda.reijnhoudt@dans.knaw.nl, andrea.scharnhorst@dans.knaw.nl*
DANS, Royal Netherlands Academy of Arts and Sciences (KNAW), the Hague, the Netherlands

[2] *rcostas@cwts.leidenuniv.nl, noyons@cwts.leidenuniv.nl*
Center for Science and Technology Studies (CWTS)-Leiden University, Leiden, the Netherlands

[3] *katy@indiana.edu*
Cyberinfrastructure for Network Science Center, School of Library and Information Science, Indiana University, Bloomington, Indiana, United States of America

Abstract
The study of science at the individual micro-level frequently requires the disambiguation of author names. The creation of author’s publication oeuvres involves matching the list of unique author names to names used in publication databases. Despite recent progress in the development of unique author identifiers, e.g., ORCID, VIVO, or DAI, author disambiguation remains a key problem when it comes to large-scale bibliometric analysis using data from multiple databases. This study introduces and validates a new methodology called seed+expand for semi-automatic bibliographic data collection for a given set of individual authors. Specifically, we identify the oeuvre of a set of Dutch full professors during the period 1980-2011. In particular, we combine author records from the National Research Information System (NARCIS) with publication records from the Web of Science. Starting with an initial list of 8,378 names, we identify ‘seed publications’ for each author using five different approaches. Subsequently, we ‘expand’ the set of publications in three different approaches. The different approaches are compared and resulting oeuvres are evaluated on precision and recall using a ‘gold standard’ dataset of authors for which verified publications in the period 2001-2010 are available.

Conference Topic
Old and New Data Sources for Scientometric Studies: Coverage, Accuracy and Reliability (Topic 2) and Management and Measurement of Bibliometric Data within Scientific Organizations (Topic 9)

Introduction

Creating correct linkages between a unique scholar authoring a work and her or his (possibly many) author name(s) is complex and unresolved. Authors might use anonymous and alias author names, names might be misspelled or change over time, e.g., due to marriage, and multiple scholars might have the very same name. Yet, science is driven by scholars, and the identification and attribution of works to individual scholars is important for understanding the emergence of new ideas, to measure the creative human capital of institutions and nations, to model the relationships and networks of researchers, and to forecast new scientific fields (Scharnhorst et al. 2012). With the 'return of the author' in bibliometrics (Scharnhorst & Garfield 2011); bibliometric indicators on the individual level (Hirsch, 2005; Costas et al, 2010; Lariviere, 2010; Vieira & Gomes, 2011); and institutional evaluation based on the individual publication output of authors over longer time periods (van Leeuwen, 2007; Zuccala et al, 2010), the ambiguity problems in allocating publications to authors have become more pressing (Costas et al, 2005, 2010). Different approaches to data collection at the individual level have been proposed in the literature (see the review by Smalheiser & Torvik, 2009), although in many cases these approaches focus on the disambiguation of author in one single database (e.g. PubMED). Recently, systems of unique author identifiers offer a practical solution, e.g. ORCID. However, they are not yet fully standardized and often rely on authors to register their own bibliographic profiles. Thus, the problem of automatically linking author names across publication, patent, or funding databases still persists.

In this paper, we present a general methodology that combines information from different data sources[157] to retrieve scientific publications covered in the Web of Science (WoS) for a given list of authors. Specifically, we trace the publications for 8,378 professors affiliated with at least one of the Dutch universities, as included in the Nederlandse Onderzoek Databank (NOD, Dutch Research Database) and displayed in the web portal NARCIS (National Academic Research and Collaborations Information System). The approach differs from prior work by the usage of an initial set of 'seed publications' for each author; and the expansion of this seed to cover the whole oeuvre of each author as represented in a large bibliographic database using an automated process. This automatic process is applied in parallel to each of the authors in the initial set. At the end, an ensemble of authors and their publications is build from the individual oeuvres. We compare five different approaches to create 'seed publications' and three approaches to 'expand' the seed. Last but not least, we assess and validate the proposed methodology against a 'gold standard' dataset of Dutch authors and their publications that was compiled by Centre for Science and Technology Studies (CWTS) and verified by the authors themselves.

[157] The idea of combining different data sources with the objective of collecting data at the individual level is not completely new (see for example D'Angelo et al, 2011) and has shown already interesting results.

The proposed methodology is able to account for different kinds of author ambiguity such as different ways of spelling a name, different ways to store a name (initials, first and last name, etc.) in different database systems, and misspellings. Homonyms, i.e., different authors with the same name, can be partially resolved using additional information about authors such as address and institution information yet, two authors might work at the same institution and on similar topics and be merged. Note that each new information source likely offers new challenges. Plus, there is much human error that is hard or impossible to detect: Mail addresses can be wrongly noted or allocated, even publications verified by the authors themselves can be wrong. Because we aim for a scalable methodology of automatic oeuvre detection we counter these ambiguities by different means:

- manual cleaning of initial sets for automatic retrieval
- manual inspection of multiple links in automatically produced mappings
- applying similarity measures (as Levenshtein distance) for string comparison
- elimination of most common names and
- elimination of publications when more than one professor from the sample with similar names are matched to the same author in a given paper (e.g. Ad de Jong and Albert de Jong assigned to the same paper with the author "A. de Jong").

The rest of the paper is organized as follows. First, a description of the different datasets is given, followed by an overview of the 'seed creation' approach applied. Second, the 'expansion of the seed' and the results and issues of performance are presented. Finally, we discuss the key results of the proposed methodology, draw conclusions, and discuss planned work.

Data

The datasets used in this study are under active development at DANS (Data Archiving and Networked Services), an institute of the Royal Netherlands Academy of Arts and Sciences (KNAW) and the Centre for Science and Technology Studies (CWTS). DANS promotes sustained access to digital research data and also provides access, via NARCIS.nl, to thousands of scientific datasets, e-publications and other research information in the Netherlands. In addition, the institute provides training and advice, and performs research into sustained access to digital information. CWTS is a centre of excellence in bibliometric analysis. It has conducted numerous bibliometric studies both for research and for evaluation, and compiled extensive data about Dutch researchers.

NARCIS/NOD database: The Dutch full professor seed

KNAW serves the NARCIS Dutch research information system (Baars et al, 2008) a web portal for a set of databases. One of them is the so-called NOD (Dutch research database) which contains information about forty thousand plus

personnel employed at Dutch research institutions (universities and other academic institutions). The person database contains metadata such as names, e-mail addresses, but also—for some scholars among them—the Dutch Digital Author Identifier (DAI). Introduced in 2008 in the Netherlands, the DAI assigns a unique identifier to every employee of a Dutch university, university of applied sciences (HBO-*Hoger beroepsonderwijs*) or research institute. Since 2006, NARCIS also harvests publications from Dutch scientific repositories. These are matched to the scholars on their DAI. A complete dump of the NARCIS database was made on April 3, 2012 and is used in this paper (Reijnhoudt et al, 2012). Specifically, we will use the set of 8,378 *hoogleraren*, or *full professors,* and their 105,128 papers to exemplify the proposed methodology. 75% of the full professors have a known DAI.

CWTS Web of Science database: High quality publication data

The in-house CWTS version of the Thomson Reuters Web of Science (WoS) consists of nearly 35 million scientific publications and hundreds of millions of citations, from 1980 up to 2012, covering all fields of science. It comprises the Science Citation Index Expanded (SCIE) as well as different enhancements made during the scientific and commercial activities of CWTS over more than 20 years. Enhancements include among others: The standardization of different fields, namely addresses, journal names, references and citation matching, and a new disciplinary classification at the paper level (Waltman & van Eck, 2012). The methodology proposed here uses the standardized address information and the new classification.

CWTS SCOPUS database: Scopus Author Identifier

Scopus is one of the largest abstract and citation databases of peer-reviewed literature. The database contains 47 million records, 70% with abstracts from more than 19,500 titles from 5,000 publishers worldwide covering the years 1996 – 2012 (http://www.info.sciverse.com/ scopus/about). Of particular interest for this study is the newly introduced 'Scopus Author Identifier' that is based on an assignment of documents to authors determined by their similarity in affiliation, publication history, subject, and co-authors (Scopus, 2009). It has been discussed that articles assigned to a particular Scopus Author Identifier tend to be articles of the author represented by that identifier, but the set of articles might be incomplete, or articles by the same author might be assigned to multiple identifiers (see Moed et al, 2012).

CWTS Gold Standard dataset: High quality publication oeuvres

Frequently CWTS' studies at the individual author level require a manual verification process in which the individual researchers check and verify their own lists of publications - for more details on this verification process see (van Leeuwen, 2007). This verification process has been applied to different sets of researchers in the Netherlands on publications from 2001 to 2010. From this

dataset of verified author-publication oeuvres we retrieve a 'gold standard' dataset by manual matching on initials, last names and organisations. This dataset consists of 1400 full professors captured in NARCIS to evaluate different author disambiguation methods. The use of a gold standard set is a common approach in bibliometric and information retrieval research (Costas & Bordons, 2008; Sladek et al, 2006).

Methodology

The main objective of this study is to develop and validate a general methodology, called seed+expand, for automatic oeuvre detection at the individual author level. Given a set of author names, we are interested to detect their publications, as many (high recall) and as correctly (high precision) as possible combining data from different databases. To exemplify and evaluate the methodology, we use the set of all 8,378 full professors included in the NARCIS database. Figure 1 shows an overview of the workflow comprising:

1. **Seed Creation**: Starting with an initial list of 8,378 full professors, we collect information on their name, affiliation, and e-mail addresses from the NOD. Next, we identify 'seed publications' in the WoS for each author using five different approaches.
2. **Seed Expansion**: Retrieval of additional papers for seed authors based on characteristics of the papers. Three different approaches are compared.
3. **Evaluation**: Results of the different seed expansion approaches are validated using standard measures for precision and recall and the CWTS 'gold standard' dataset of authors for which verified publications in the period 2001-2010 are available.

All three parts are detailed subsequently.

Seed Creation

The first step of the methodology consists of the creation of a reliable 'seed' of publications for the 8,378 target professors. An element of this set consists of a triplet of elements (publication identifier, person identifier and author position in the paper). The identifiers come from different databases; and the author position indicates if the scholar is first, second, or *n*th author for that publication. The accuracy of the seed is very important as the precision and recall of the final oeuvre detection will be significantly higher if the precision of the seed is high. It is important to bear in mind that during the expansion phase of the methodology it will not be possible to add papers for those professors that are not already in the seed. So, for this phase in the methodology it is not the recall on papers but the recall on authors that matters. Five different approaches of creating 'accurate' seeds are explored here:

- E-mail seed (EM): A seed based on the matching of the e-mail of the professor with the publications in Web of Science.

- Three author-address approaches (RP, DL, AL): These seeds are based on different combinations of the name of the professors and the affiliation(s) of that professor matched with the Web of Science.
- DAI seed (DAI): This approach builds upon the publications in NARCIS that have been attached to the professors through the Dutch Author Identifier.

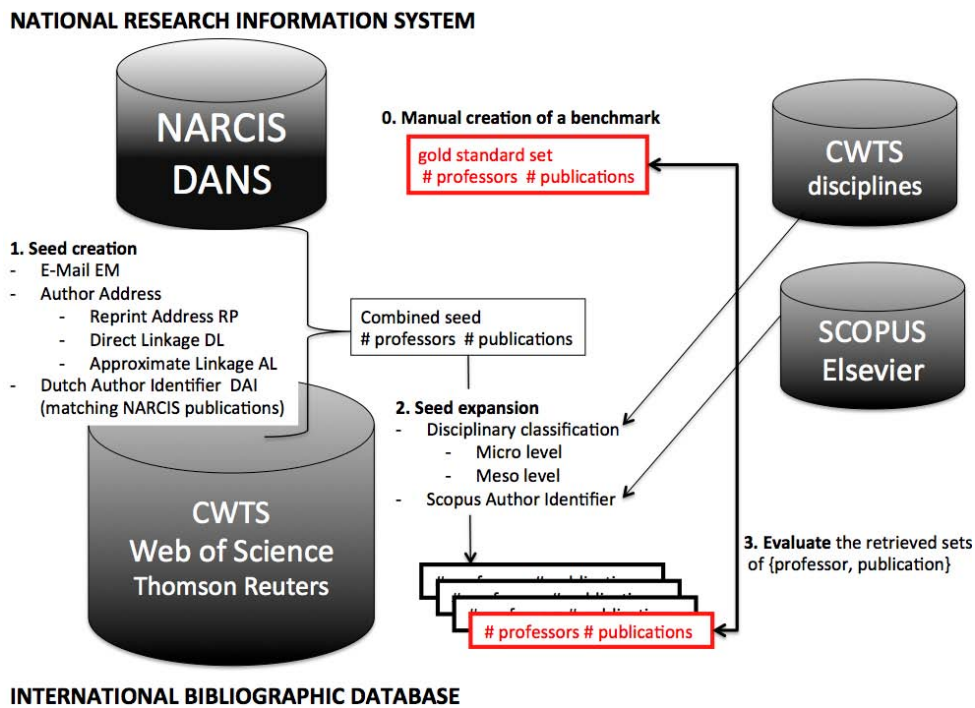


Figure 1. General workflow and relevant data sources

E-mail Seed (EM)

In the NOD system, e-mail addresses are attached to scholars directly or via their affiliations. Hereby, e-mail addresses of the target professors are simply matched against e-mail addresses of authors found in the papers in Web of Science. This approach produced a seed for 4,786 different authors (57% of the professors from our list) with at least one paper found in WoS, see also Table 1. As e-mail addresses are uniquely attached to one scholar[158] and are seldom transferred to other scholars[159], this approach is assumed to be most accurate.

[158] Exceptions exist with addresses like info@ or dep@

[159] We expect sometimes an e-mail is transferred to another researcher with the same (or very similar) name in the same organization when the previous e-mail holder has left the organization.

Three Author-Address approaches

Three approaches combine author names and affiliation data in the NOD system and match them to WoS affiliation data to retrieve relevant publications. This approach was only feasible thanks to standardization of WoS affiliations and addresses by CWTS. However, some parts of this task also required manual handling and checking. As a result of this, 92% of the papers with a Dutch organization in the WoS have a matched counterpart in the NOD organizations. These are the only 652,978 papers that can be considered for the Author-Address approach seeds as described in the following paragraphs.

Reprint author (RP): In scientific publications, the reprint address refers to the address of the corresponding author in charge of managing requests that a publication may generate. In the WoS database this reprint address appears directly linked to the author, thus offering a direct and "safe" connection between an author and an organization that can be directly extracted from the publication. Thus, the creation of this seed consists of the matching of the name of the professor and his/her affiliation as is recorded in the NOD with the reprint author name and the reprint affiliation.

Direct linkage author-addresses (DL): 69% of publications WoS include data on the linkage between the authors and their organizations as they appear in the original publications. For instance, if the original publication featured three authors and two organizations this linkage of authors and organizations is indicated as follows (Figure 2):

Author A(1)(2), Author B(2), Author C(1)
Organization G, Organization H

Figure 2. Example of direct author organization linkage

Indicating that Author A is linked to Organization G and H; Author B is linked to Organization H, and Author C is linked to Organization G. As in the RP-based approach, the names and the affiliations of the professors are matched with the author-affiliation linkages of the publications, detecting those publications that, based on this author-affiliation linkage, could belong to the target professors.

Approximate linkage author-addresses (AL): The other 31% of publications did not have a direct linkage recorded in the database. Thus, authors and affiliations of the publications were recorded, but there was no way to tell which author is affiliated with what organization. This approach detects as seed publications of the target professor those publications that share the same name and affiliation as the professor in the same record. The AL approach has the potential problem of wrongly attributing a publication to a target professor if the name of the author

and the institute both appear on a paper. For instance, referring back to Figure 2, if a homonym of 'Author B' (i.e., another scholar with the same name) appears in a paper where 'Organization H' also appears, the real 'Author B' might get this paper wrongly attributed to him/her.

DAI Seed (DAI)

This seed creation approach starts with publications that are in the NARCIS database attributed by means of the DAI to the target professors. However, these publications are not necessarily WoS publications (there may be books, theses, or journal articles not covered in the WoS). For this reason, it was necessary to perform a matching process between the bibliographic records for the professors with a DAI extracted from NARCIS and the WoS database. The NARCIS papers were matched with the WoS publications on journal, year, title, and first page. This way, we were able to create a new seed, based on the publications covered in the Web of Science that were also in the NARCIS database for the target professors.

Combining the seeds

Table 1 shows the resulting numbers of publications and professors created by the five different seed creation approaches. Publications are counted once per seed, even if they appear several times for different professors. The last column shows the number of professors that were found exclusively by this particular seed method. So, if the AL approach would not be used, the number of professors found would drop only by 76, whereas not using the EM approach would result in a drop of 790 professors. At the end, all seed results are combined (added) and cleaned for duplicates leaving us with 6,989 unique professors and corresponding 174,568 publications.

Table 1. Result sets obtained by different seeds

Seed Method	*CWTS Publications*	*NARCIS Full Professors*	*Full Professors Unique to This Seed*
EM	40,826	4,786	790
RP	81,079	5,819	149
DL	79,515	5,749	158
AL	28,837	5,018	76
DAI	30,322	2,742	162
Total unique in combined seed	**174,568**	**6,989**	

To further improve seed quality, we remove multiple assignments and common names. Multiple assignments refer to the cases where more than one professor is matched to the same paper, with the same author position number. Clearly this is wrong, as only one researcher should be matched to one author. In order to keep

the seed as precise as possible and thus sacrificing some recall for precision, all these records have been removed (see Table 2, 'remove multiple assignments'). The top 5% most common author names (first initial-last name pairs) from the former two seed-approaches (RP and DL) and the top 10% from the least precise approach (AL), thus trying to keep the level of 'noise' (i.e., false positives) in the seed to a minimum. (See Table 2, column 'remove common names').
The resulting seed comprises 6,753 professors (80% of initial set of 8,378) with 157,343 unique papers.

Table 2. Pruning the seeds to increase precision

Seed Method	*Number of Found Professors*	*Remove Multiple Assignments*	*Remove Common Names*
EM	4,786	4,786	4,786
RP	5,819	5,696	4,648
DL	5,749	5,629	4,675
AL	5,018	4,864	3,147
DAI	2,742	2,742	2,740
Total unique in combined seed	**6,989**	**6,947**	**6,753**

Seed Expansion

In this second phase, we use the 6,753 author profiles and associated papers in the seed to identify additional publications by these authors in the WoS database. Three approaches have been explored and are detailed subsequently.

Two CWTS Paper-Based Classifications (Meso and Micro)

These two approaches use a new paper-based classification that has been developed at CWTS (Waltman & van Eck, 2012) based on the citation relationships of individual publications. It has been applied to publications between 2001 and 2011, excluding the Arts and Humanities. The hierarchical classification has three levels, with a medium-level classification that comprises 672 'meso-disciplines', and a lower-level classification that includes more than 20,000 different 'micro-disciplines.' We assume that within small disciplinary clusters (meso and micro) there is a rather low probability that two professors share the same name. Hence, we assign all papers within a meso/micro-discipline that have the same author names to one professor. Incorrect assignments might occur when two persons with the same name work in the same subfield.
Performing assignments at the meso-discipline level results in an increase of 34% to 211,202 unique papers, the micro-disciplines yield a subset thereof with 194,257 unique papers, an increase of 23%.

Scopus Author Identifier Approach

A third approach to expand the publications of professors is to use one of the already existing author identifiers. We choose the Scopus Author Identifier because it has been introduced for all authors in the Scopus database. Here, the 157,343 WoS publications from our initial seed were matched to Scopus publications and their 6,753 authors were matched with Scopus authors to derive their Scopus author identifier. As shown in Figure 3, 614 WoS seed authors had no Scopus author identifier; 2,977 authors had exactly one Scopus author identifier; and all others had more than one. All Scopus author identifiers were used to retrieve additional Scopus publications that were traced back to WoS publications via bibliographic matching on journal, title, etc. The resulting set has 266,105 unique papers, an increase by 69%—the largest number of all three approaches.

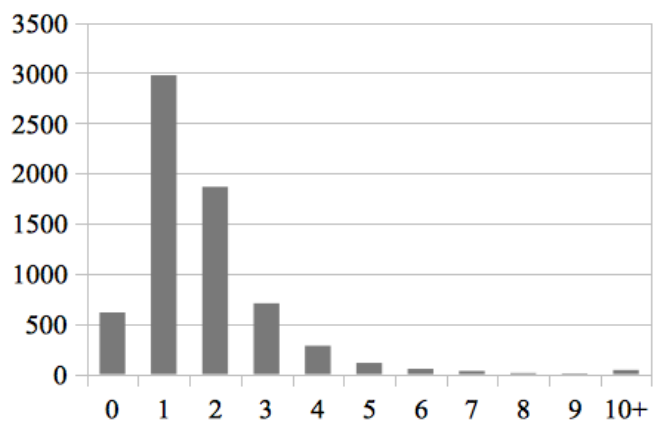


Figure 3. Number of authors (y-axis) from the seed with a given number of matched Scopus author identifiers (x-axis). The 47 authors with more than 10 were ultimately discarded.

Evaluation

To evaluate the three seed expansion approaches, their result sets are compared to the CWTS gold standard dataset introduced in section 2.4. The expansion of the seed by the different approaches has been performed on the whole WoS (from 1980 to 2011). But to evaluate the approaches we restrain the result of the expansion to publications published between 2001 and 2010, the same time period as the gold standard set. Exactly 1,400 of the 6,753 authors (21%) are in the gold standard dataset - only 63 professors are not accounted for. These 1,400 authors and their 57,775 associated papers will be used to measure precision and recall achieved by the different approaches.

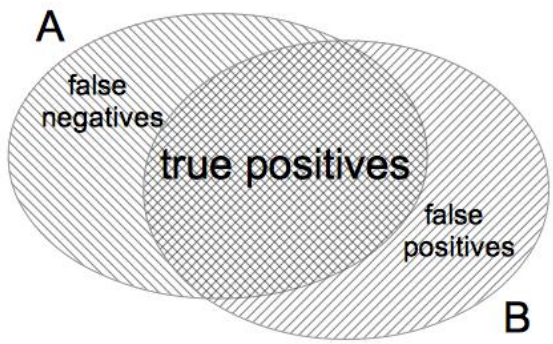


Figure 4. Gold standard set (A) versus the result of the expansion (B)

Precision and recall are widely used to measure how well an information retrieval process performs. Precision is defined as the retrieved relevant records (true positives) divided by all retrieved records (both true and false positives). Recall on the other hand is the number of retrieved relevant records divided by the number of records that should have been retrieved (the true positives and the false negatives). Thus, we can score the performance of our different approaches according to these two parameters, see Table 3

Column 2-4 in Table 3 present the three approaches individually. In general, the values for recall are equally high for the three approaches. Regarding the precision there is a slight difference. As expected, the micro-discipline set has a higher precision and lower recall than the meso-disciplines approach. The Scopus author identifier approach, with 63,460 professor-paper combinations and a precision of 87.3, ends up exactly in between.

Table 3. Performance of the three expansion approaches - individually and combined

	Scopus Identifier	*Meso*	*Micro*	*ScopusI & Meso*	*ScopusI & Micro*
True pos. **(A∩B)**	55,405	55,459	55,394	55,509	55,460
False pos. **(¬A∩B)**	8,055	10,430	7,212	13,200	10,260
False neg. **(A∩¬B)**	2,370	2,316	2,381	2,260	2,315
Precision	87.3	84.2	88.5	80.8	84.4
Recall	95.9	96.0	95.9	96.1	96.0

The last two columns show the combination of two approaches: *Scopus author id plus meso-disciplines* and *Scopus author id plus micro-disciplines*. As can be expected, the recall increases, whereas the precision declines. The increase in the recall is rather small, and the number of false negatives is high. This indicates that both approaches miss roughly the same papers. Apparently there are some

publications in the oeuvres of some researchers that are hard to find using the kind of approaches presented in this paper.

Conclusions

Of the 8,378 professors in our target list we identified at least one publication for 6,753 (80%) professors, which gave us a seed into their oeuvre (as far as covered by the WoS). Combining all publications of all professors we started with a set of 6,753 authors and 157,343 unique publications. After the expansion of the individual publication seeds with the Scopus author id approach and the micro-disciplines approach we find the same recall on the gold standard dataset, and a comparable precision, as shown in Table 3. The Scopus author id approach finds more unique papers (266,105 vs. 194,257). This can be attributed to the restrictions on the disciplines classification on period (2001-2011) and subject, e.g., the Arts and Humanities WoS publications are not included (Waltman & van Eck, 2012).

The gold standard dataset used for evaluation covers 1,400 (or 21%) of the 6,753 professors and we assume the precision and recall results can be extrapolated to the entire data collection. It is important to remark that the results of precision are slightly conservative due to the fact that for some authors some publications were still missing in their verified set of publications. This happens particularly with authors with high numbers of publications, in fact Smalheiser & Torvik (2009) indicated that this happens when authors have more than 300 publications. In other words, although the precision of our gold standard set is 100% (basically we can assume that all are correct publications, as verified by their authors) it seems that the recall of the golden standard set is not necessary 100%. Thus, the values of 'wrong' publications obtained through our methodologies might not be as high in reality, and thus we can consider this measure to be the upper bound of false positives that we could expect for the whole analysis, because true values will likely be smaller. The methodology developed in this paper will be further applied in an impact study of the set of Dutch full professors, retrieving citations to all their publications. We would like to point out that our methodology, relying on domain specific scholarly communication, is sensitive towards the disciplinary composition of the author set, e.g., authors that publish mostly books are underrepresented. This will be explored in further analysis.

Note that the success of cross-database retrieval and author disambiguation heavily depends on access policies of the hosting institutions, and the quality of the databases involved. Even if access is given, extensive institutional collaboration is required to interlink and harmonize databases. Initiatives such as ORCID (Foley & Kochalk, 2010) with the idea of a central registry of unique identifiers for individual researchers or bottom-up networked approaches such as the VIVO international researcher network (Börner et al, 2012) that assigns unique VIVO identifiers to each scholar, aim to provide processes and data structures to assign and keep track of unique scholars and their continuously evolving oeuvres. The data collected by ORCID and VIVO can be used as

additional 'gold standards' in future evaluation studies. The methodology presented here can be applied to retrieve publications for scholars with a valid ORCID and VIVO from the existing commercial and public data sources. Ultimately, unique author identifiers are required for the comprehensive analysis of science, e.g., using altmetrics (Wouters & Costas, 2012), and also for models of science (Scharnhorst et al. 2012) using data from multiple databases.

Acknowledgements

The authors acknowledge detailed comments and suggestions provided by Vincent Lariviere and Kevin W. Boyak for an early version of this manuscript. Part of this work was funded by the NoE European InterNet Science (EINS) (EC Grant 288021) and the National Institutes of Health under award U01 GM098959.

References

Baars, C.; Dijk, E.; Hogenaar, A.; van Meel, M. (2008). Creating an Academic Information Domain: A Dutch Example. In: Bosnjak A., Stempfhuber M. (Eds.) Get the Good Current Research Information System (CRIS) Going: Ensuring Quality of Service for the User in the European Research Area. Proceedings of the 9th International Conference on Current Research Information Systems, Maribor, Slovenia, June 05-07, pages 77-87 (Online at http://depot.knaw.nl/5628/).

Börner, K.; Ding, Y.; Conlon, M.; Corson-Rikert, J. (2012). *VIVO: A Semantic Approach to Scholarly Networking and Discovery*. San Rafael, Calif.: Morgan & Claypool Publishers.

Costas, R.; Bordons, M. (2005). Bibliometric indicators at the micro-level: some results in the area of natural resources at the Spanish CSIC. *Research Evaluation*, 14(2): 110-120.

Costas, R.; Bordons, M. (2008). Development of a thematic filter for the bibliometric delimitation on interdisciplinary area: The case of Marine Science. *Revista Española de Documentación Científica*. 31(2): 261-272.

Costas, R.; van Leeuwen, T.N.; Bordons, M. (2010). A bibliometric classificatory approach for the study and assessment of research performance at the individual level: The effects of age on productivity and impact. *Journal of the American Society for Information Science and Technology*, 61(8): 1564-1581.

D'Angelo, C.A.; Guiffrida, C.; Abramo, G. (2011). A heuristic approach to author name disambiguation in bibliometrics databases for large-scale research assessments. *Journal of the American Society for Information Science and Technology*, 62(2): 257-269.

Foley, M.J.; Kochalk, D.L. (2010). Open Researcher and Contributor Identification (ORCID). Proceedings of the Charleston Library Conference <http://dx.doi.org/10.5703/1288284314850>.

Moed, H.F.; Aisati, M.; Plume, A. (2012). Studying scientific migration in Scopus. *Scientometrics*, Online at http://link.springer.com/content/pdf/10.1007%2Fs11192-012-0783-9

Lariviere, V. (2010). A bibliometric analysis of Quebec's PhD students'contribution to the advancement of knowledge. PhD Thesis. Montreal: McGill University.

Reijnhoudt, L.; Stamper, M.J.; Börner, K.; Baars, C.; Scharnhorst, A. (2012) NARCIS: Network of Experts and Knowledge Organizations in the Netherlands. Poster presented at the Third annual VIVO conference, August 22 - 24, 2012 Florida, USA, http://vivoweb.org/conference2012. Online at http://cns.iu.edu/research/2012_NARCIS.pdf

Scharnhorst, A., Garfield, E. (2011) Tracing Scientific Influence. *Dynamics of Socio-Economic Systems*, 2 (11): 1–31. Preprint online at http://arxiv.org/abs/1010.3525

Scharnhorst, A.; Börner, K. and Van den Besselaar, P. eds. (2012) *Models of Science Dynamics*. Berlin, Heidelberg: Springer Berlin Heidelberg. doi:10.1007/978-3-642-23068-4.

Scopus (2009). Frequently Asked Questions – Author Identifier. Online at http://www.info.sciverse.com/documents/files/scopus-training/resourcelibrary/pdf/FAQ_Author_Identifier_09.pdf

Sladek, R.; Tieman, J.; Fazekas, B.S.; Abernethy, A.P.; Currow, D.C. (2006). Development of a subject search filter to find information relevant to palliative care in the general medical literature. *Journal of the Medical Library Association.* 94(4): 394-401.

Smalheiser, N.R., Torvik, V.I. (2009). Author name disambiguation. *Annual Review of Information Science and Technology*. 43: 287-313.

Van Leeuwen, T.N. (2007). Modelling of bibliometric approaches and importance of output verification in research performance assessment. *Research Evaluation*, 16 (2): 93-105.

Vieira, E.S.; Gomes, J.A.N.F. (2011). An impact indicator for researchers. *Scientometrics*, 89 (2): 607-629.

Waltman, L.; van Eck, N.J. (2012). A new methodology for constructing a publication-level classification system of science. *Journal of the American Society for Information Science and Technology*, 63(12): 2378-2392.

Zuccala, A.; Costas, R.; van Leeuwen, T.N. (2010). Evaluating research departments using individual level bibliometrics. *Eleventh International Conference on Science and Technology Indicators*. Leiden: CWTS-Leiden University.

THE SHORTFALL IN COVERAGE OF COUNTRIES' PAPERS IN THE SOCIAL SCIENCES CITATION INDEX COMPARED WITH THE SCIENCE CITATION INDEX

Grant Lewison[1] and Philip Roe[2]

[1] *grantlewison@aol.co.uk*
King's College London: Research Oncology, Guy's Hospital, Great Maze Pond, London SE1 6RT (UK)

[2] *philip@evaluametrics.co.uk*
Evaluametrics Ltd, 157 Verulam Road, St Albans, AL3 4DW (UK)

Abstract
Most physical science research papers are universal in their interest and are published in international journals in English. However much social sciences research is of primary interest to readers in the authors' country and so is often published in the national language and in journals not processed for the SSCI. We wondered if it was possible to estimate the shortfall in coverage of this research by the SSCI by comparing the ratio of papers from a given country in the SSCI only to the numbers in both the SCI and SSCI with the corresponding ratio for the USA, with account taken of the relative expenditures on social sciences and medical research. For this purpose, we examined sets of papers with each of a selected title word chosen from one of four subject categories, and found, as expected, that Anglophone countries showed much less shortfall than the large continental European ones (France, Germany, Italy, Spain) or those in east Asia (China, Japan, South Korea). The data in the paper can be used to estimate how many social sciences papers are likely to be "missing" from the SSCI for the leading countries.

Conference Topics
Scientometric Indicators: Relevance to Social Sciences (Topic 1); Old and New Data Sources for Scientometric Studies: Coverage, Accuracy and Reliability (Topic 2)

Introduction
Much bibliometric work is concerned to describe the evolution of scientific subject areas and to identify the leading actors – countries, institutions and individuals. In the physical sciences, this task has been made easier in recent years with the advent of the Web of Science (WoS) and SCOPUS, which provide nearly comprehensive coverage of journals in all fields of science, most of which are in English. The WoS in particular has increased its coverage recently of journals published other than in western Europe or north America. For example, Indian cancer research coverage in the WoS included only 9 Indian journals in 1995 but as many as 35 in 2010 (Lewison and Roe, 2012) The decline in

coverage of Indian journals in the 1990s by the Science Citation Index (SCI) was the main reason for the apparent decline in output of Indian scientists (Basu, 1999). Relatively small increases or declines in national scientific output can be important matters for politicians in government or opposition, respectively. So, since these indicators may depend critically on the coverage of national journals by the WoS, it is important to investigate any shortfall in coverage.

What do we mean by "shortfall"? This means that the number of papers from a country in the SSCI is less than its total output in peer-reviewed social sciences journals because the database coverage is less comprehensive than it is for social sciences journals that are mainly used by US scientists, *i.e.* that there is a national bias in the journal selection process used by Thomson Reuters. This process may not be value-free: Klein and Chang (2004) suggested that there was "a variety of evidence of bias in favor of journals of a social democratic orientation and against journals of a classical liberal orientation". In particular, journals not in English are poorly represented in the SSCI compared with their abundance in Ulrich's Periodicals Directory, a list of over 300,000 periodicals first compiled in 1932 by Carolyn Ulrich, the head of periodicals at the New York Public Library (Wiegand, 1990). This is now a commercial database, and is coupled with the Ulrich's Serials Analysis System (Jacsó, 2012). Between 1992 and the present, on average more than 94% of the SSCI papers (all document types) were in English.

To investigate the extent of the shortfall in representation, we made the initial assumption that the SSCI coverage of US output would be reasonably comprehensive, and therefore that its ratio of papers in the SSCI only to papers in both databases would be a value that would be the norm for the given subject area. (This may well be conservative because, as we shall see, the USA spends rather less on social science research relative to that on medical research compared with other countries.) We then determined this ratio for some 30 leading countries, and divided it by the ratio for the USA to give a "shortfall ratio" (SR) which would depend on the country and also on the subject matter. For example, Table 1 gives the data for a title word (properly, stem) "adolescen" for the USA and two other countries, one of which (Australia) has a much smaller shortfall than does France, *i.e.*, the SR is closer to unity. The numbers are the integer counts of papers from the respective countries.

Table 1. Example of calculation of shortfall ratio for Australia and for France for the title stem, "adolescen".

Country	*ISO*	*SCI+SSCI*	*SSCI only*	*SSCI/both, %*	*Shortfall ratio*
United States	US	7388	2948	39.9	
Australia	AU	825	239	29.0	0.73
France	FR	416	59	14.2	0.36

There are likely to be two reasons for a shortfall in papers in the SSCI – in the above table, 27% for Australia and 64% for France. One is the paucity of SSCI coverage of the journals from a given country that might be thought worthy of

inclusion, but the second may be more important, namely that the country actually does less social sciences research in comparison with its physical sciences research than the USA does in the same subject areas. If this is so, then the real shortfall in SSCI coverage is correspondingly less than would appear from calculations such as those whose results are shown in Table 1. Effectively, the measured shortfall above is the product of two factors: the under-representation of a country's social sciences journals and the under-performance of its social sciences research. We can readily determine the value of SR for a given subject area, but we need to calculate the under-representation of social sciences journals from data on social sciences research expenditures compared with physical sciences ones.

Table 2. Title words selected for the study, with the numbers of Canadian SSCI papers containing each of them in 2009-11 (articles and reviews only, and not in SCI). HUM = human behaviour, MED = medical, ORG = organisational, PSY = psychological.

HUM words	*CA*	*MED words*	*CA*	*ORG words*	*CA*	*PSY words*	*CA*
adolescen	331	cancer	104	activity	234	brain	233
adult	290	clinical	106	assessment	212	cognitive	207
children	865	disorder	363	decision	188	dement	31
community	421	injury	81	education	309	depression	92
family	297	medical	48	evaluation	217	learning	387
health	735	nursing	53	evidence	466	memory	167
human	186	pain	59	impact	390	mental	192
older	244	patients	116	information	227	phobia	12
population	141	randomized	52	knowledge	225	psycholog	256
risk	426	treatment	208	management	277	schizophrenia	21
women	453	trial	82	outcomes	195	stress	103
young	200			performance	320	visual	82
				quality	184		
				survey	306		
Mean:	**382**		**116**		**268**		**149**

Methodology

Selection of subject areas.

We chose four main areas: human behaviour (designated HUM), medical (MED), organisational (ORG) and psychological (PSY). For each of these, we selected about a dozen different title words (or stems, such as "adolescen" in Table 1 to cover both adolescence and adolescent(s)). They were chosen from the list of the title words most frequently used in Canadian social sciences papers in the three years 2009-11 that were in the SSCI but not in the SCI. Canada was chosen because it was expected that its social sciences papers would be well covered in the SSCI and that they would encompass a wide range of topics. The selected

words and stems were as listed in Table 2, with the numbers of Canadian papers in the SSCI only that contained them in their titles. It appears that there were many more papers in human behaviour (HUM) and organisational studies (ORG) than in psychology (PSY) and medicine (MED) – the latter papers would have been largely in the SCI.

Searches on the Web of Science and analysis of SR values

For each title word, we counted the numbers of papers in the SCI plus the SSCI, and in the SSCI only, excluding papers in the SCI. We then calculated, for each country and for each word in the above table, the ratio of these two numbers, and divided this by the corresponding ratio for the USA to give a nominal shortfall value. For each country, we then calculated the mean of this value for each of the four sets of title words, together with its standard deviation, from which we determined the standard error of the mean (s.e.m.). For this purpose we used the standard WoS software, supplemented by some of our own.

OECD data on expenditures on social sciences research and R&D personnel

We obtained data on research expenditures and numbers of research scientists in major fields from the Science and Technology Indicators published annually by the Organisation for Economic Co-operation and Development (OECD) in Paris. It was difficult to make standard comparisons for all the 31 countries (including the USA) as many countries did not provide the OECD with statistics for each year, and some gave no breakdown by field. We selected expenditures (in millions of constant 2005 US dollars) and numbers of research personnel for medical & health and for social sciences. However, for some countries and/or years, social sciences were combined with humanities, and we had to estimate the share of this total that represented social sciences. The intention of this data-gathering exercise was to estimate the effort (expenditure and personnel) devoted by the individual countries to social sciences research relative to that given to medical & health research. However, the years of the data inevitably varied from country to country, and we needed to assume that this ratio was reasonably constant. We attempted to obtain data for 2006, on the grounds that expenditures then would be likely to generate research papers three or four years later, but such data were not available for all countries. If they were not, we used data from the latest year for which they were published.

Data from Ulrich's Periodicals Directory

The purpose of this exercise was to obtain data on the numbers of social science journals in the individual countries that claimed to use peer-review for comparison with the numbers that were covered in the SSCI. (For the latter count, we excluded journals in other subject areas that had a few social science papers.) However, Ulrich's does not distinguish between social sciences and humanities, so we needed to inspect the titles of all the journals in this category in order to count the number that corresponded in their apparent coverage to the

journals processed for the SSCI. These included archaeology, economics and business, education, international relations, law and sociology, but not history (except for the history of science and medicine). Allocation of Ulrich's journals either to social sciences or to humanities was a nice decision in many instances, and was carried out by PR. The intention was to use these data as a simple check on the results of the calculation of SSCI shortfall, which would be based on the apparent shortfall and the difference between the ratio of social sciences research to medical & health research in the country and in the USA. Because the SSCI selects journals on the basis of their contribution to the international literature, the shortfall based purely on the numbers of journals in the Ulrich's directory might be quite inaccurate, but at least the value based on the Ulrich's data should bear some resemblance to the truth, and it would serve to give some credibility to the calculated value of the shortfall of SSCI papers.

In order to see if these shortfall data agreed with those in the Ulrich's Periodicals Directory, we first determined the identity and country of publication of all the journals covered in the SSCI during 2009-11. Of course, the shortfall will depend not only on the number of journals covered in the SSCI but also the number of articles in each one. Also, researchers from a given country may publish their social science papers in journals published in other countries – and indeed, the country of publication may not be obvious to intending authors who may only be aware of the nationality of the editor(s) and will probably submit their mss electronically. However, we have assumed that there will be a tendency for social scientists to seek a national readership for their research, and so may publish some papers in national journals listed in Ulrich but not processed for the SSCI. A further complicating factor is that many non-Anglophone countries (notably the Netherlands and Germany) publish many of their journals in English in order to attract a wider range of contributors and readers.

The listing of all the SSCI journals, with their countries of publication and languages, is a non-trivial task, even with the excellent search facilities available on the WoS[160]. Individual papers have the publisher's address and language, but it is not possible currently to search the WoS by country of publication. This means that the bibliographic details of large numbers of selected papers have to be downloaded and analysed in order to list journals with their countries of publication. We compared these data with the numbers of peer-reviewed social science journals covered by Ulrich from the different countries.

[160] It is possible to list the names of all the journals whose papers are included in a given search, but the names do not always correspond to the correct names of the journals – hyphens are replaced by spaces, and ampersands are omitted, which makes matching difficult. Moreover, many journals are not published from the country that is given in their title.

Results

Apparent shortfall of papers in SSCI.

The primary results of the study are shown in Table 3 for the 30 leading countries (the USA is omitted as the values are unity, by definition). In the table, SR values with an s.e.m. of less than 10% are printed in **bold**, ones with s.e.m. less than 20% printed in normal type, and ones with s.e.m > 20% printed in *italics*.

Table 3. Apparent shortfall in papers in SSCI only (SR), based on ratio of SSCI-only papers to ones in both SCI and SSCI compared with corresponding ratio for the USA, 2009-11. HUM = human behaviour, MED = medical, ORG = organisational, PSY = psychological.

Country	*ISO*	*ALL*	*HUM*	*MED*	*ORG*	*PSY*
Israel	IL	0.96	**0.93**	0.85	**1.03**	1.03
New Zealand	NZ	0.96	0.89	0.72	**1.08**	1.14
Australia	AU	0.95	**0.90**	**0.91**	**0.98**	0.99
Norway	NO	0.93	**0.74**	1.03	**0.89**	1.07
Canada	CA	0.90	**0.92**	**0.91**	**0.81**	**0.96**
England	EN	0.88	**0.91**	**0.81**	**0.94**	**0.89**
Scotland	SC	0.77	0.83	0.65	**0.73**	0.88
Sweden	SE	0.76	0.56	0.88	0.66	*0.95*
Spain	ES	0.75	0.64	0.64	**0.75**	0.95
Romania	RO	0.75	0.69	*0.99*	0.56	*0.76*
Netherlands	NL	0.74	**0.70**	**0.58**	**0.84**	0.83
Finland	FI	0.71	**0.49**	*0.86*	0.78	0.71
Ireland	IE	0.68	**0.80**	0.60	**0.78**	*0.52*
Malaysia	MY	0.65	0.56	*0.27*	0.96	*0.82*
Germany	DE	0.64	**0.53**	0.55	**0.60**	0.87
Singapore	SG	0.63	0.65	*0.56*	0.82	0.50
Belgium	BE	0.61	**0.56**	0.46	**0.59**	0.83
Switzerland	CH	0.57	**0.45**	0.50	**0.54**	*0.80*
Denmark	DK	0.57	0.36	*0.64*	0.59	0.68
Taiwan	TW	0.55	**0.47**	*0.31*	0.91	0.52
Brazil	BR	0.55	*0.52*	0.58	0.55	0.54
Russia	RU	0.44	0.47	*0.30*	*0.48*	*0.50*
Greece	GR	0.44	0.34	*0.39*	0.50	0.52
France	FR	0.40	0.33	*0.40*	0.37	*0.48*
China	CN	0.37	0.35	*0.37*	0.47	0.27
South Korea	KR	0.36	0.29	*0.33*	0.47	0.36
Italy	IT	0.35	**0.30**	*0.32*	0.34	*0.45*
Hungary	HU	0.34	0.33	*0.26*	0.41	*0.34*
Japan	JP	0.27	0.22	*0.33*	0.28	*0.23*
Poland	PL	0.23	0.21	*0.19*	0.25	0.28
Mean Values			**0.57**	**0.57**	**0.66**	**0.69**

The countries are ordered on the basis of apparent shortfall, with those countries showing the least difference from the USA at the top, and the ones showing the greatest difference at the bottom. It is not surprising that the countries at the top are, in the main, Anglophone, whose papers are most likely to be published in international journals and in English, *i.e*., the ones that stand the best chance of being selected for inclusion in the SSCI. However, there are a few anomalous placings – Norway is very high, but Ireland, Malaysia and Singapore are only in the middle of the table. The shortfall appears to be greater for human behaviour and medicine (mean value of SR = 0.57) than for organisational matters (SR = 0.66) and for psychology (SR = 0.69). A few of the SR values exceed unity: this does not necessarily mean that the shortfall is negative, but rather that the country involved devotes relatively more effort to social sciences research compared with the type of work that also appears in the SCI than the USA does. The amount of relative effort on different science fields is discussed in the next section.

Relative effort on social science research and on medical & health research.

The results taken from the OECD science & technology indicators are shown in Tables 4 and 5. The first of these shows the financial expenditures on medical & health research and for social sciences research in the given year (data for some of the countries listed in Table 3 are not available), and the ratio between them. This varies from 0.82 (Russia, where there is very little biomedical research) to 0.17 (USA and China, which appear to do little social sciences research in comparison with their biomedical work, Sarewitz, 2013). The data for the USA are quite old (1998 is the latest year for which they are available), and in the quinquennium after that year the budget of the National Institutes of Health (NIH) doubled in real terms (Greenberg, 1999; Check, 2002), so it is unlikely that the ratio has become larger. However, since 2003, the NIH budget has barely increased because of US budgetary difficulties (Rosbach, 2011).

Table 4. Expenditures on medical & health research and on social sciences research in 23 leading countries in the given year, USD2005 million, constant prices and PPP. Ratio is soc sci / (medical & health + soc sci)

Code	*Year*	*M & H*	*Soc sci*	*Ratio*		*Code*	*Year*	*M & H*	*Soc sci*	*Ratio*
RU	2007	43	200	0.82		BE	2006	410	238	0.37
IL	1999	189	347	0.65		CA	2005	1822	987	0.35
ES	2007	623	992	0.61		NL	2001	739	397	0.35
PL	2006	117	136	0.54		KR	1998	313	155	0.33
FI	2006	265	231	0.47		SE	2007	777	329	0.30
IE	2006	104	88	0.46		IT	1987	799	324	0.29
JP	2001	4483	3645	0.45		CH	2000	263	106	0.29
DK	2006	369	217	0.37		DE	2000	2463	863	0.26
NO	2007	427	250	0.37		SG	2005	421	96	0.19
TW	2006	399	233	0.37		US	1998	8731	1841	0.17
AU	2006	1322	768	0.37		CN	2000	170	34	0.17

The second table shows the numbers of research personnel. Far fewer countries report these numbers; most of the totals given here are the sum of the numbers in three sectors: government, higher education and private-non-profit. The correlation between the ratios obtained from expenditure and personnel data is just positive but very modest, see Figure 1. However, if the spot for the most extreme outlier (PL, Poland) is removed, then the correlation for a linear trend-line improves from $r^2 = 0.09$ to 0.27, and if the spot for IT, Italy, is also removed then $r^2 = 0.41$.

Table 5. Numbers of research personnel in medical & health research and in social sciences research in 11 leading countries in the given year. Ratio is soc sci / (medical & health + soc sci)

Code	*Year*	*M & H*	*Soc sci*	*Ratio*		*Code*	*Year*	*M & H*	*Soc sci*	*Ratio*
IL	2008	661	1524	0.70		NL	2009	12668	7884	0.38
IE	2006	1042	1121	0.52		CH	2000	2650	1494	0.36
IT	2009	23175	21786	0.48		JP	2003	78688	41343	0.34
AU	2006	18012	16211	0.47		DK	2006	4996	2463	0.33
ES	2009	28646	22660	0.44		PL	2006	3323	834	0.20
SE	2003	5327	3634	0.41						

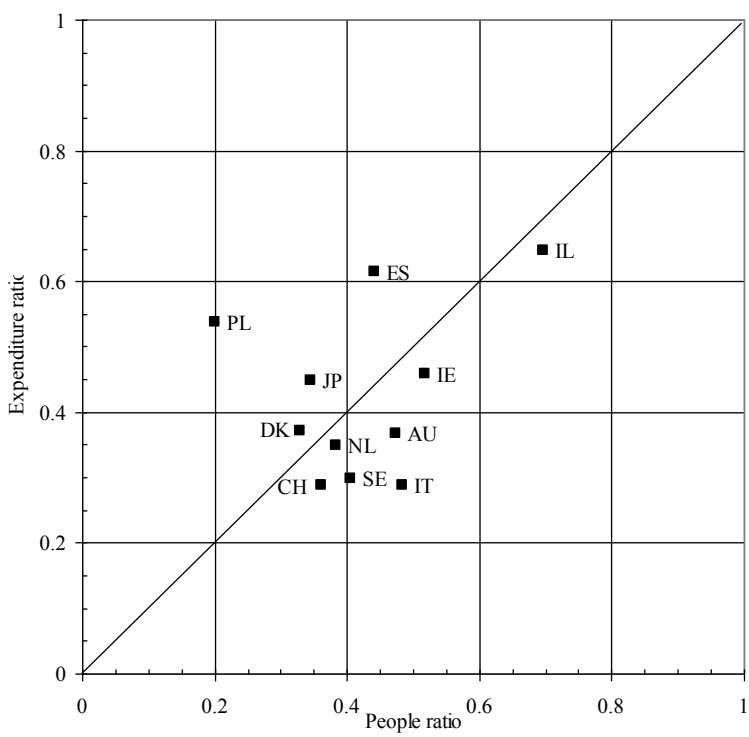


Figure 1. Ratio of social science to medical & health plus social science expenditures compared to the ratio for people, for 11 OECD countries, various dates.

It is not clear which of the two data sets is the more reliable, so for countries where both indicators are available, a mean value has been taken to be applied to the SR values in Table 3. Because the USA does relatively little social sciences research in comparison with its activity in medical & health research, most other countries have a higher ratio, and would therefore be expected to generate relatively more papers in the SSCI than they appear to do in Table 3. This means that the shortfall in SSCI coverage is even greater than suggested by this table, and the best estimates for the shortfall, based on multiplication of the SR values by the ratio of the ratio for the country to the ratio for the USA, are those shown in Table 6, where the calculations are displayed. Values are only available for the 27 countries for which an estimate of relative effort on social sciences to (medical & health plus social sciences) is available; this excludes France, New Zealand and the UK.

This last table shows that the shortfall in SSCI coverage is slightly more than half for Canada and Norway, around two thirds for some northern European countries (Germany, Sweden, Netherlands) and Australia, about three quarters for some others (Switzerland, Denmark, Finland, Ireland, Belgium), and more than 90% for Poland and Russia.

Table 6. Calculated shortfall in papers in SSCI only (SR), based on ratio of SSCI-only papers to ones in both SCI and SSCI compared with corresponding ratio for the USA, 2009-11, and corrected for difference in relative effort on social science research compared with medical & health research from OECD S&T data.

Country	*ISO*	*HUM*	*MED*	*SS/(M&H+SS)*	*SR* HUM*	*SR* MED*
Canada	CA	0.92	0.91	0.35	0.45	0.44
Norway	NO	0.74	1.03	0.37	0.34	0.47
Australia	AU	0.90	0.91	0.42	0.37	0.37
China	CN	0.35	0.37	0.17	0.35	0.37
Germany	DE	0.53	0.55	0.26	0.35	0.36
Sweden	SE	0.56	0.88	0.35	0.27	0.43
Netherlands	NL	0.70	0.58	0.37	0.32	0.27
Switzerland	CH	0.45	0.50	0.32	0.24	0.26
Denmark	DK	0.36	0.64	0.35	0.18	0.31
Finland	FI	0.49	0.86	0.47	0.18	0.31
Ireland	IE	0.80	0.60	0.49	0.28	0.21
Belgium	BE	0.56	0.46	0.37	0.26	0.21
Israel	IL	0.93	0.85	0.67	0.24	0.22
Spain	ES	0.64	0.64	0.53	0.21	0.21
Taiwan	TW	0.47	0.31	0.37	0.22	0.14
South Korea	KR	0.29	0.33	0.33	0.15	0.17
Italy	IT	0.30	0.32	0.39	0.13	0.14
Japan	JP	0.22	0.33	0.40	0.09	0.14
Poland	PL	0.21	0.19	0.37	0.10	0.09
Russia	RU	0.47	0.30	0.82	0.10	0.06

Table 7 lists the numbers of journals published by the top ten countries that we were able to identify in the SSCI; it is clear that the distribution is highly skewed, with 47% coming from Canada and the USA and 46% from Europe.

Table 7. Numbers of SSCI journals published in 10 leading countries present in 2009-11

Country	*ISO*	*Number*	*%*		*Country*	*ISO*	*Number*	*%*
United States	US	1211	46.6		Australia	AU	35	1.3
United Kingdom	UK	744	28.6		France	FR	21	0.8
Netherlands	NL	159	6.1		Brazil	BR	20	0.8
Germany	DE	110	4.2		Canada	CA	19	0.7
Spain	ES	48	1.8		Switzerland	CH	18	0.7

It proved difficult to list the journals in the Ulrich list that mapped clearly onto the fields covered by the SSCI journals so that a comparison could be made. For several countries, the number of "relevant" Ulrich journals was actually smaller than the number of journals processed for the SSCI. For example, there were 110 SSCI journals published in Germany (68 in English, 40 in German) but the tally of "relevant" journals in Ulrich was only 62: 43 in German and 29 in English. This suggests that almost all the German-language journals were covered in the SSCI, so there should have been relatively little shortfall for German authors, but this is unlikely to be the case.

Table 8. Numbers of papers (articles and reviews) in the SSCI but not in the SCI, 2009-11, with the number in own country language(s) and percentage.

ISO	*SSCI x SCI*	*Own lang.*	*% OL*		*ISO*	*SSCI x SCI*	*Own lang.*	*% OL*
BR	4361	2655	60.9		NO	3404	54	1.6
RU	1396	765	54.8		NL	10707	166	1.6
ES	10450	3627	34.7		SE	4939	60	1.2
DE	14755	3403	23.1		TW	4549	15	0.3
FR	6987	1337	19.1		CN	7450	10	0.1
PL	1214	217	17.9		DK	2518	1	0.0
CH	3945	433	11.0		IL	3653	0	0.0
HU	622	58	9.3		RO	990	0	0.0
JP	3492	186	5.3		FI	2509	0	0.0
BE	4093	183	4.5		GR	1418	0	0.0
IT	5683	178	3.1		KR	3285	0	0.0

Finally, we give data on the numbers and percentages of papers from the non-Anglophone countries that appeared in the SSCI but not the SCI and were in the national language(s). This gives some indication of the possibilities for social scientists to publish in national journals, despite the pressures in many countries for publications in English in order to gain a wider readership. The list is, perhaps surprisingly, headed by Brazil (Portuguese) and Russia. Some countries have no

national-language journals in the SSCI, and this may account for their relatively smaller shortfall seen in Table 3 – for example, Israel, Romania and Finland.

Discussion

It was clear that something was amiss with those countries for which the number of "relevant" Ulrich journals was smaller than or only slightly above the number of SSCI journals: this was the case for Germany, Italy, Poland, Spain and Switzerland within Europe, and also South Korea and Taiwan. But for the other nine countries, there was a positive correlation between the SR values in Table 6 and the ratio between SSCI journal numbers and Ulrich "relevant" numbers, see Figure 2.

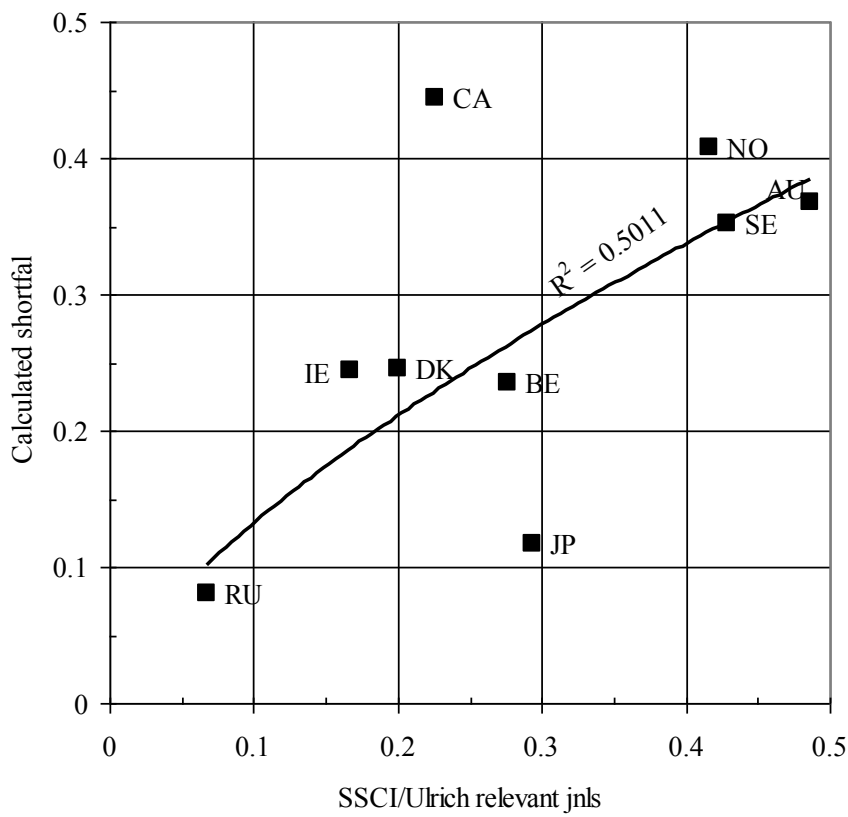


Figure 2. Scatter plot of the calculated shortfall values for social science papers (from Table 6) for nine countries compared with the ratio of numbers of their journals in the SSCI compared with those in Ulrich's Periodicals Directory for social science peer-reviewed academic journals.

The correlation is moderate, and would be better except for two outliers: Canada, whose calculated shortfall value is greater than expected, and Japan, whose value is smaller. It is possible that Ulrich's directory does not adequately cover all the Japanese journals: there were only 17 social science ones listed, half the number published in Brazil.

The results presented here can only be regarded as rather approximate, and additional work is clearly needed to refine them. However, it is apparent that the shortfall is real and quite large, and biggest for Russia, Poland and Japan; somewhat smaller for Italy, Spain and Belgium; less again for the Scandinavian countries; and least for the Anglophone countries (Australia, Canada, the UK), as would be expected.

Acknowledgments

The authors gratefully acknowledge the assistance of Lauren Hutchinson with the downloading of the data from the WoS on country outputs.

References

Basu, A. (1999) Science publication indicators for India: Questions of interpretation. *Scientometrics*, 44, 347-360.

Check, E (2002) Bush's budget boost puts NIH on target for doubled figures *Nature*, 415, 459.

Greenberg, DS (1999) Washington – a big budget and big changes on the way for US NIH *The Lancet*, 354, 1621.

Jacsó P (2012) Analysis of the Ulrich's Serials Analysis System from the perspective of journal coverage by academic databases *Online Information Review*, 36, 307-319

Klein DB & Chiang E (2004) The Social Science Citation Index: A black box – with an ideological bias? *Econ Journal Watch*, 1, 134-165

Lewison, G. & Roe, P. (2012) The evaluation of Indian cancer research. *Scientometrics*, 93, 167-181.

Rosbash, M (2011) A threat to medical innovation *Science* 333, 136.

Sarewitz, D (2013) Science must be seen to bridge the political divide. *Nature*, 493, 7.

Wiegand WA (1990) Carolyn Ulrich *Supplement to the Dictionary of American Library Biography*. Libraries Unlimited, Englewood, Colorado, 136-168.

SOCIAL DYNAMICS OF RESEARCH COLLABORATION: NORMS, PRACTICES, AND ETHICAL ISSUES IN DETERMINING CO-AUTHORSHIP RIGHTS (RIP)

Barry Bozeman[1], Monica Gaughan[2] and Jan Youtie[3]

[1] *bbozeman@uga.edu;*
University of Georgia, Department of Public Administration and Policy, 201 Baldwin Hall, Athens, Georgia 30602-1615 (USA)

[2] *gaughan@uga.edu*
University of Georgia, Department of Health Policy and Management, College of Public Health, Athens, Georgia 30602 (USA)

[3] *jan.youtie@innovate.gatech.edu*
Georgia Institute of Technology, Enterprise Innovation Institute, 75 Fifth Street, Suite 300, Atlanta, Georgia 30308 (USA)

Abstract
As co-authorship has become common practice in most science and engineering disciplines and, with the growth of co-authoring has come a fragmentation of norms and practices, some of them discipline-based, some institution-based. It becomes increasingly important to understand the practices, in part to reduce the likelihood of misunderstanding in collaborations among authors from different disciplines and fields. Moreover, there is also evidence of widespread satisfaction with collaborative and co-authoring experience. In some cases the dissatisfactions are more in the realm of bruised feelings and miscommunication but in others there is clear exploitation and even legal disputes about, for example, intellectual property. Our paper is part of a multiyear study funded by the U.S. National Science Foundation and draws its data from a representative national survey of academic scientists working in Carnegie Extensive ("Research I") universities (n=641). The paper tests hypotheses about the determinants of collaboration effectiveness.

Conference Topic
Collaboration Studies and Network Analysis (Topic 6) and Sociological and Philosophical Issues and Applications (Topic 13).

Introduction
Of late there has been a growing concern about "contributorship" and, particularly, that authors may be included as co-authors in research in which they had no research role and, worse, may not even understand. Contributorship is defined as authors declaring in detail, in advance of publication, their individual contributions to scholarly papers (Rennie 2000, p. 1274). Contributorship

policies and norms are viewed as increasing transparency and fairness; The impact of institutionalized standards is a timely aspect of the proposed work. Increasingly, journals and professional associations are adopting rules and guidelines about "contributorship" (Rennie, 1998). While journals and professional associations continue to make much needed contributorship policies, it remains the case that these policies are based more on anecdotal information and even rumours than on empirical research. Our paper seeks to provide such an empirical basis.

Related Studies of Research Collaboration Dynamics

During the past decade or so, researchers, especially those in the biomedical sciences (e.g. Rennie et al., 2000; Cohen 2004), have begun to focus on ethical issues and the "dark side" of collaboration. Lagnado (2003) argues that trust in the meaning of co-authorship has eroded. Levsky and colleagues (2007) describe potentially troubling trends in authorship in medical journals between 1995 to 2005, including honorary authorship, ghost authorship, duplicate and redundant publications and most important, authors' refusal to accept responsibility for their articles despite their readiness to accept credit for professional purposes. They note that causes of the trends continue to be unknown but that the relationship between authorship and career pressures on academic physicians are clear.

Outside of biomedical fields, research on the ethics and socio-political dynamics of scientific collaboration (Shrum, et al., 2001, 2007) remains scarce. Perhaps this scarcity is owing to the view (we think mistaken) that such problems are neither as pervasive nor as troublesome in other STEM fields. To be sure, biomedical research is different. In most fields of science, technology, engineering and mathematics (STEM) there is little potential for unethical behavior to affect clinical trials (Devine 2005; Klingensmith and Anderson 2006) and there are no pharmaceutical industry representatives providing services as "phantom" co-authors. Nonetheless, our preliminary studies (Bozeman, et al., 2012) show that many of the same ethical threats and problems documented in biomedical fields occur in other STEM fields, albeit with somewhat different causes and impacts.

Far from being restricted to biomedical fields, problems in scientific collaboration are ubiquitous in science. Some of these problems are ethical (Shrum et al., 2001), others practical (Bozeman and Corley, 2004; Lee and Bozeman, 2005), some pertain to collaboration among individuals (Katz and Martin, 1997; Bozeman and Corley, 2004), and some to collaboration among institutions (Chompalov and Shrum, 1999). The literature on scientific collaboration not only identifies problems in collaboration but also possible solutions. For example, Marusic (2004) and Pichini (2005) describe the many international Uniform Requirements for coauthorship information and the complex but poorly understood relationship between contributorship and grants, promotion, and admittance to professional associations. Most work is case-based or anecdotal and, as a result, neither the scientific community nor policy-makers have much

systematic, empirically based evidence of the possible pitfalls of collaboration and contributorship.

Our study draws from the abundant literature on scientific collaboration (see Katz and Martin, 1997 and Bozeman, Fay and Slade, 2013 for overviews), especially questions associated with scholarly manuscript authorship (Tulandi et al. 2008; Chompalov, et al., 2002) to analyze the ethical challenges for participants in collaborative STEM research. By developing from the authors themselves information about collaboration dynamics, norms and social and ethical dilemmas, this work provides insights into the potential of new public policies and designs that will promote effective collaboration. Using a web-based survey, we seek to develop strong empirical knowledge of STEM researchers' norms, behaviours, and perceptions about collaboration and co-authoring, especially the assigning of credit.

The foundation of STEM research-based knowledge is peer-reviewed publication of research findings. Due to increasingly interdisciplinary work and large-scale, the assignment of authorship for publication is complex and sometimes confusing. Allocation of credit and responsibility for authorship is an important issue and it must be resolved if STEM research results are to be managed effectively (Devine 2005).

While there are many problems with co-authoring credit and contributorship, some are well known and familiar. One problem is that scientific fields and even work groups within fields vary substantially in their practices for assigning co-authoring credit. In some cases first authorship means that the individual made the most significant scientific and intellectual contributions to the research, but in other cases it means that the individual was the lab director or the principle investigator and may have had little or no direct involvement in the research (Mowatt et al. 2002). An alternative practice – alphabetizing authorship order – would presume to reflect more "fairness" but it also lacks explicit information as to which author is primarily responsible for the work. The decision about assigning credit is highly varied and often provides only an oblique signal as to who has done what. But a decision prior to co-authorship status (who is a co-author and what it means) is the problem of how co-authorship issues are decided. As decision analysts have known for years, often process is the primary determinant of outcome (Brockner and Wiesenfeld, 1996). While there is remarkably little evidence about collaboration and co-authorship decision processes and norms, most agree that these vital processes affect not only scientific career trajectories and advancement but very course of science (Katz and Martin, 1999; Melin, 2000). The choice of scientific topics and the configuration of research teams depend in part on collaborative and co-authorship norms. Researchers have considerable autonomy in their collaboration choices and collaboration strategies are based in part of judgments about the conferring of co-authorship and status (Heffner, 1981; Bozeman and Corley, 2004). The issue is who decides.

Some attribute problems with sorting out contributorship to the explosion in research and the funding imperatives driving collaboration among investigators from multiple sites and numerous disciplines (Devine et al. 2005, Drenth 1998). Ultimately the system of scientific authorship is built on trust that the published work reflects the data and analysis of the authors (Lagnado 2003). We contend that co-authorship choices and perceptions relate closely to the integrity of the research. Specifically, the integrity of the research can be undermined in (at least) the following ways. In cases where authors, especially lead or corresponding authors, make no substantial contribution to the research, authorship claims are essentially scientific fraud. Control of authorship award and credit can in some instances be a "weapon" that the powerful use to obtain resources, knowledge or obedience from the less powerful. Since grants and contracts and other important research-related resources are provided in part on the basis of scientific reputation and apparent productivity, measured in terms of quantitative and quality of publication, misattribution of co-author credit undermines the effective allocation of science resources. In cases where "legitimate" collaborators are not included as co-authors and not provided their due, there is a possibility that data privileges and other resource controls can act as, essentially, a restraint of trade. Conflicts of interest in STEM research, often revealed in authorship of scholarly literature, are problematic for obvious reasons (McCrary et al. 2000). While our study cannot deal with the full range of ethical problems and implications flowing from co-authorship and collaboration issues, we can begin to prepare the empirical basis for understanding these problems. Absent more detailed knowledge of norms, practices and perceptions about collaboration, it is difficult to even begin to understand the extent of ethical hazard. According to Rennie (2000, p. 91), "the general consensus appears to be that identifying and publishing specific contributions of authors is a venture that shows promise. But its utility must be demonstrated." Our study seeks to assess the processes, dynamics, and utility of various approaches to contributorship decisionmaking.

The Research Focus and Hypotheses

The focus of this work is on data from a web survey. This paper examines determinants of predictors of collaboration experiences in the "most recent co-authored research publication." We used this wording approach in the survey to prevent the respondents from having to provide a specific citation, thereby threatening their anonymity. Here we examine difficulties related to (1) persons not being credited whom are perceived as deserving credit, (2) persons being credited who were perceived as not deserving, (3) gender-based conflict. Our central hypothesis is that undesirable collaboration outcomes are associated with collaborations in which co-authorship credit is never discussed explicitly. Other hypotheses associate undesirable collaboration outcomes with large collaborations, author collaboration motivations, negative past experiences, mixed gender co-author groups, and collaborator geographic distance. Controls for author's PhD award year and gender were included.

Data

The analysis is based on a web survey of 641 non-medical academic researchers in science and technology disciplines in US doctoral research universities (Carnegie Doctoral/Research Universities—High). A sampling frame of science and technology fields was developed using NSF's categories in its Survey of Earned Doctorates. Health sciences was excluded (because of its medical orientation) while economics was added to incorporate social science practices into the survey. The resulting frame was based on 14 disciplines in biology, chemistry, computer science, mathematics, engineering, and economics. The sampling frame called for one male and one female faculty member from each randomly selected department at a given university because qualitative interviews suggested that gender would be a significant factor; in the event that no female faculty members were affiliated with the department, two male researchers were selected. The target sampling frame, which assumed a 50% response rate, resulted in 2,996 faculty and another 216 postdocs. We were able to collect contact information for 2,574 individuals in the sampling frame; of these 2,189 were of sufficient quality as indicated by an electronic mail verification software program. Pilot surveys performed in April and May of 2012 used 400 of these, leaving 1,789 for the final survey. Six waves of survey invitations and reminders were sent in October and November of 2012. One percent were not at their office location, while another 5% explicitly opted-out of participation. In all, we received 641 completed or mostly completed online questionnaires, for a 36% response rate. Respondents were very similar to the population in terms of gender, rank and departmental discipline. Given that we oversampled females and certain departments, results are re-weighted to reflect the population distribution as indicated in the NSF Survey of Doctorate Recipients 2006 (the most recently available survey).

Results

Undesirable collaboration outcomes are measured with respect to the most recent co-authored research publication. A logit model (Table 1) is based on a variable "problems," indicating lack of disagreement with statements about denial of deserved co-authorship, receipt of undeserved co-authorship, and gender-based conflict based on responses to a 10-point scale from strongly agree to strongly disagree. The number of co-authors in the recent paper is logged ("lnrecent_numcoauth"). Whether this publication involved explicit discussions about co-authoring credit ("creditdis_yesno") addresses the main hypothesis. Descriptive statistics indicate that 40% of respondents engaged in explicit discussions about co-authoring credit. We captured the percentage of co-authors of the paper that are male "permale" and the percent at other universities than those of the respondent "perothuniv." Importance of motivations for collaborations to increase research productivity and help a co-author's career ("mov_productivity" and "mov_helpcoauthor") are represented in responses to a 10-point scale ranging from not important at all to extremely important. The

existence of past negative behaviors in terms of a co-author having made no contribution to the research is included as well ("careerbad_undeservedcoaur"). Controls for the year of the author's PhD ("yearphdr") and whether or not the author is male or female (male2) are included.

Explicit discussion of co-authorship credit is negatively associated with the likelihood of problems. The number of co-authors increases the likelihood of problems, which was in the expected direction. However, there is an inverse relationship between the percentage of authors at another university and the likelihood of problems, which is counter to our geographic hypothesis. Authors with a bad undeserved co-authorship experience in their past were more apt to report problems in their recent research publication, although motivations to collaborate were not significant. The year a PhD was granted is positively associated with problems, indicating that younger academics are more likely to experience problems than are their older counterparts. Although gender overall was not significant, interactions males in mathematics and computer science and males in engineering reduced the likelihood of problems. The model is statistically significant and 80% of responses are correctly classified.

Table 1. Logit model of Likelihood of Problems in Recent Research Publication

Variables	*Logit(problems)*	*Robust standard errors (parens.)*
creditdis_yesno	-0.624	(0.288)**
lnrecent_numcoauth	0.712	(0.184)***
permale	-0.003	(0.004)
perothuniv	-0.011	(0.004)**
mov_helpcoauthor	0.043	(0.042)
mov_productivity	-0.012	(0.057)
careerbad_undeservedcoaur	1.108	(0.276)***
yearphdr	0.026	(0.013)*
male2	0.709	(0.562)
male2bio	-0.721	(0.673)
male2phys	-0.933	(0.688)
male2math	-1.446	(0.741)*
male2eng	-1.007	(0.605)*
Constant	-53.561	(26.256)**

Log likelihood=-254.84898, Wald (chi square) significant at 1%, Pseudo R2=.13
* significant at 10%; ** significant at 5%; *** significant at 1%

Summary

There are many conceptual studies and case studies of the social and organizational processes by which researchers make decisions about contributions, credit sharing and authorship shared credit (see for example Fine and Kurdek, 1999). However , systematic, large sample, studies remain scarce (e.g. Vinkler, 1993; Floyd, et al., 1994), hence this work's contribution.

Acknowledgments

The authors thank Derrick Anderson and Daniel Fay for their assistance. This study was undertaken with support from the National Science Foundation under Award # 1026231. Any opinions, findings, and conclusions are those of the authors and do not necessarily reflect the views of the sponsors.

References

Bozeman, B. & E. Corley. (2004). Scientists' collaboration strategies: implications for scientific and technical human capital. *Research Policy* 33:599-616.

Bozeman, B. & M. Gaughan. (2007). Impacts of grants and contracts on researchers' interactions with industry. *Research Policy,* 33, 5: 694-707.

Bozeman, B., J. Youtie & D. Libaers (2006). Institutionalization of university research centers: The case of the National Cooperative Program in Infertility Research. *Technovation* 26: 1055-163.

Bozeman, B., J. Youtie, C. Slade, & M. Gaughan. (2012). Nightmare Collaborations, Paper presented at the Annual Meeting, International Society for the Social Study of Science, Cophenhagen, 2012.

Bozeman, B., D. Fay & C. Slade (2013). Research Collaboration and Academic Entrepreneurship: A State of the Art Review. *Journal of Technology Transfer,* in press, on line first.

Brockner J. & B. Wiesenfeld (1996). An integrative framework for explaining reactions to decisions: Interactive effects of outcomes and procedures. *Psychological Bulletin* 120(2): 189-208.

Chompalov, I., J. Genuth, & W. Shrum. (2002). The organization of scientific collaborations." *Research Policy* 31:749-767.

Chompalov, I. & W. Shrum. (1999). Institutional collaboration in science: A typology of technological practice." *Science Technology & Human Values* 24:338-372.

Cohen, J. J. (2004). Realizing our quest for meaning. *Academic Medicine: Journal of The Association of American Medical Colleges, 79*(5), 464-468.

Devine, E. B., Beney, J., & Bero, L. A. (2005). Equity, Accountability, Transparency: Implementation of the Contributorship Concept in a Multi-site Study. *American Journal of Pharmaceutical Education, 69*, 455-459.

Drenth, J. P. H. (1998). Multiple Authorship: The Contribution of Senior Authors. *JAMA: Journal of the American Medical Association, 280*(3), 219-221.

Fine, M. A. & Kurdek, L. A. (1993) Reflections on Determining Authorship Credit and Authorship Order on Faculty-Student Collaborations, *American Psychologist*, 48, 11, 1141-1152.

Finholt, T. (2002). Collaboratories. *Annual Review of Information Science and Technology*, 36, 73-108.

Heffner, A.G., (1981). Funded research, multiple authorship, and subauthorship collaboration in four disciplines. *Scientometrics* 3, 5–12.

Katz, J. (1994). Geographical proximity and scientific collaboration, *Scientometrics*, 31(1): 31-43

Katz, J. S. & B. R. Martin. (1997). What is research collaboration? *Research Policy* 26:1-18.

Klingensmith, M. E., & Anderson, K. D. (2006). Educational scholarship as a route to academic promotion: a depiction of surgical education scholars. *American Journal of Surgery, 191*, 533-537.

Lagnado, M. (2003). Increasing the trust in scientific authorship. *British Journal of Psychiatry* 183(1), 3-4.

Lee, S. & B. Bozeman. (2005). The Effects of Scientific Collaboration on Productivity. *Social Studies of Science, 35, 5:673-702*.

Levsky, M. E., Rosin, A., Coon, T. P., Enslow, W. L., & Miller, M. A. (2007). A Descriptive Analysis of Authorship Within Medical Journals, 1995--2005. *Southern Medical Journal,* 100: 371-375.

McCrary, S. V., Anderson, C. B., Jakovljevic, J., Khan, T., McCullough, L. B., Wray, N. P. (2000). A National Survey Of Policies on Disclosure of Conflicts of Interest in Biomedical Research. *New England Journal of Medicine*, 343(22): 1621-1626.

Marusic, M., Bozikov, J., Katavic, V., Hren, D., Kljakovic-Gaspic, M., and Marusic, A. (2004). Authorship in A Small Medical Journal: A Study of Contributorship Statements by Corresponding Authors. *Science and Engineering Ethics*, 10(3): 493-502.

Melin, G., (2000). Pragmatism and self-organization: research collaboration on the individual level, *Research Policy* 29, 3: 1140-1670.

Mowatt, G., Shirran, L., Grimshaw, J. M., Rennie, D., Flanagin, A., Yank, V., et al. (2002). Prevalence of Honorary and Ghost Authorship in Cochrane Reviews. *JAMA: Journal of the American Medical Association, 287*(21): 2769-2771.

Pichini, S., Pulido, M., and Garcia-Algar, O. (2005). Authorship in Manuscripts Submitted to Biomedical Journals: An author's Position and Its Value. *Science and Engineering Ethics*, 11(2): 173-175.

Rennie, D. (1998). Freedom and Responsibility in Medical Publication: Setting the Balance Right. *JAMA: Journal of the American Medical Association, 280*(3), 300-302.

Rennie, D., & Flanagin, A. (1994). Authorship! Authorship! *JAMA: Journal of the American Medical Association*, 469.

Rennie, D., Flanagin, A., & Yank, V. (2000). The Contributions of Authors. *JAMA: Journal of the American Medical Association*, 89.

Rennie, D., Yank, V., & Emanuel, L. (1997). When authorship fails. A proposal to make

Shrum, Wesley, Ivan Chompalov, & Joel Genuth. 2001. Trust, Conflict and Performance in Scientific Collaborations. *Social Studies of Science* 31: 681-697.

Shrum, Wesley, Joel Genuth & Ivan Chompalov (2007) *Structures of Scientific Collaboration*. Cambridge, MA: MIT Press.
Singer, N. (2009). Medical papers by ghostwriters pushed therapy, *New York Times,* August 4, 21-22
Tulandi, T., Elder, K., & Cohen, J. (2008). Responsibility and accountability of authors and co-authors. *Reproductive BioMedicine Online*, 763-764.

SOFTWARE PATENTING IN ASIA

Poh-Kam Wong[1] and Yuen-Ping Ho[2]

[1] *pohkam@nus.edu.sg*
Entrepreneurship Centre, National University of Singapore, 21 Heng Mui Keng Terrace, Level 5, Singapore 119260

2 yuenping@nus.edu.sg
Entrepreneurship Centre, National University of Singapore, 21 Heng Mui Keng Terrace, Level 5, Singapore 119260

Abstract
This paper examines software patents trends in Asia using patents granted to Asian inventors by the USPTO from 1980 to 2011. The various definitions of software adopted in prior literature are summarized and two classification-based definitions are used in the analysis. We found that globally, software patenting has grown faster than other types of USPTO-granted patents, especially more so over the last decade. Two thirds of software patents are invented in the major software producing economies of North America, Germany, France and the United Kingdom. One quarter of software patents are invented in Asia, with Japan accounting for much of this share. Excluding Japan, Asia contributes a more modest 7%. However, software patenting in non-Japan Asia is far outpacing the rest of the world over the last decade, even growing faster than in the major software producing economies of North America and Europe. This is driven by rapid growth in several key economies, namely Korea, India and Taiwan, with China close behind. Nonetheless, the contribution of software to national patents portfolios of non-Japan Asian economies is still relatively low compared to the global average. The exception is India, where software patents account for 38% of patents in the last 5 years.

Conference Topic
Technology and Innovation Including Patent Analysis (Topic 5)

Introduction
Patents as a protection mechanism for software is a relatively recent phenomenon dating to the 1990s, when court decisions in the US widened the scope of patentable subject matter to include computer-implemented methods and processes. There remains considerable disagreement in the international community on whether it is desirable to allow software to be patented rather than relying on the copyright protection mechanism. Much of the disagreement stems from differing views on the macroeconomic impact of software patents, particularly the effects of a patents regime on the level of innovation in software development (Jaffe and Lerner, 2006). The debate remains largely unresolved as there has been little empirical research on this topic (Bessen, 2011).

At the firm level however, there are a number of persuasive arguments favouring patents over traditional copyright for software IP, the most important being the stronger protection afforded by patent laws. With the growing ubiquity of digital technologies, companies that invest heavily in software development are seeking more concrete ways to safeguard their software inventions and advance their commercial interests. This has sparked a growing trend of registering software IP through the patent system, which provides for stronger protection and more clear-cut litigation against infringements. Empirical research shows that there has been a dramatic increase in the propensity to patent software since the pivotal US court decisions on the patentability of computer programs (Bessen and Hunt, 2007).

However, the literature has not specifically focused on software patents originating from Asia. Hence, despite the well-documented economic strength of Asian economies and the prominence of Asian companies in the information and communications technology (ICT) sector, little is known of the contribution of Asian inventors to the rising trend of software patenting. This paper is a first attempt to address this gap. Specifically, we seek to ascertain if the global growth in software patents has been mirrored in Asia, and the relative importance of software in the patents portfolios of Asian economies. We also examine differences between the various Asian economies to identify the regional leaders and laggards in software patents. While prior research examined software patenting up to the mid-2000s only, our analysis extends to 2011. This is important, as the explosion of software patenting occurred only over the last decade.

Table 1 Global Computer Software Spending

	Global Spending on Computer Software		
	Total USD bil	% change y.o.y.	Share of Software in ICT Spending (%)
2003	199		8.43
2004	230	15.6	8.66
2005	253	10.0	8.86
2006	275	8.7	8.94
2007	296	7.6	8.91
2008	312	5.4	8.82
2009	305	(2.2)	8.97
2010	325	6.6	9.01
2011	357	9.9	8.71
Average Growth 2006-2011	5.4%		

Sources: OECD Information Technology Outlook 2010; UNCTAD Information Economy Report 2012

Recent Trends in the Global Software Market

In this section, we present an overview of the global software market to provide additional context for our analysis. As shown in **Table 1**, worldwide spending on computer software has increased substantially from USD 199 billion in 2004, to USD 357 billion in 2011.[161] Software accounts for 9% of total spending on ICT, a share that has held steady over the years. Excepting 2009, when the global economy was reeling from the financial crisis of 2007-08, spending on computer software has grown annually at rates in excess of 5%, and in excess of 10% in the earlier part of the post millennium decade. Between 2006 and 2011, computer software spending grew at an average 5.4% annually

Table 2 Software Spending in Asia and Selected Advanced Economies, 2011

	Software Spending 2011 (USD billion)	Share of Economy/ Region in World Total Software Spending	Software Spending Intensity (Share of Software in Total ICT Spending)
ASIA	**46**	**12.9**	**4.0**
Japan	15	4.1	4.1
China	19	5.2	4.4
Hong Kong	0.5	0.2	2.7
Taiwan	1.5	0.4	5.8
India	2.3	0.6	2.4
South Korea	2.9	0.8	3.3
Singapore	1.2	0.3	8.6
Other Asia	4.4	1.2	4.0
NORTH AMERICA	**151**	**42.3**	**11.7**
United States of America	138	38.8	12.2
EU / EFTA	**138**	**38.6**	**12.2**
France	17	4.9	11.0
Germany	24	6.8	11.2
United Kingdom	24	6.7	12.9
World Total	357	100.0	8.7

Source: UNCTAD Information Economy Report 2012

The vast share of software spending takes place in North America and Europe, as shown in **Table 2**. Asian economies account for 12.9% of global spending on

[161] It is noted that data on software sales and spending do not quite capture the full size of software activities, as many companies do in-house software development. Also the figures may not fully include IT services such as contract programming, nor software which is embodied in other technology products.

software, led by Japan and China who contribute 4.1% and 5.2% respectively of the world total. We also observe that there is relatively low software spending intensity among individual Asian economies. Software spending intensity is proxied by the share of software in the economy's total ICT spending. This averaged 4% across Asia, and ranged from 2.7% in Hong Kong to 8.6% in Singapore, which is the only Asian economy with software spending intensity approaching the global average of 8.7%. Comparably, we observe that software spending contributes more than 10% of total ICT spending in advanced economies such as USA, France and Germany.

Having established an understanding of global trends in software spending, we seek to ascertain if these patterns are echoed in patenting trends. Is the production of software patents keeping pace with the expansion of the global software market? There appears to be relatively low expenditure on software in Asia compared to North American and Europe, with Asian economies concentrating higher proportions of spending on the hardware and communications segments of ICT. Does software patenting in Asia similarly lag behind the more advanced economies?

Data and Methods

Source for Patents Data

Patent laws and examination standards vary by jurisdiction. While an ever-present factor in comparative patents analysis, this issue gains increased significance when examining patents in the field of software. There are vastly divergent standards for granting software patents in the different national patent granting offices, with some offices adopting more restrictive stances than others. As a result of several key judgments by the US Supreme Court and the specialist Federal Circuit Court, the US Patents and Trademark Office (USPTO) now has among the broadest and most inclusive guidelines for patentability of software. Currently, a *practical application* of a computer-related invention is patentable by the USPTO, as are business methods. Comparably, the European Patent Office (EPO) has stringent standards for allowing patents to be granted for software inventions. The European Patent Convention (EPC) explicitly states the exclusion of “programs for computers” from patentability, although case law shows several exceptions to this standard. Among the Asian economies, standards for software patentability also vary considerably. Along with the UPSTO, the Japanese Patent Office is regarded as having more lenient laws pertaining to software patents. In India, patent laws are almost as stringent as the EPC, limiting patentability to a computer program in the form of its “technical application to industry”. Patent laws in Korea, Taiwan and China fall between the USPTO's liberal standards and the stricter provisions of the EPO.

The use of a common patent-granting agency eliminates concerns pertaining to differing standards of examination in a cross-border context. All analysis in this paper is based on patents granted by the USPTO. The USPTO is selected as one of the largest patents granting offices in the world and because the USA is the premier target market for technology products and services. This is even more so for software applications in both commercial and consumer domains. With the USPTO's liberal guidelines for patentability of computer programs and computer-implemented methods, inventors and companies from around the world also have additional incentive to seek software patent protection in the USA.

Patent counts and patents data from 1980 to 2011 were extracted from the Patsnap website, which provides an online database of USPTO patents and a web-based search engine. Of the extracted detailed fields, the inventor's country of residence is used to assign the nationality of a patent. A patent is categorized as an Asian patent if at least one of its inventors is residing in an Asian economy. With this convention, a patent with multiple inventors from multiple economies will be assigned multiple countries of origin and will be included in the patent counts of each country.

Organization of Data to Reflect Developments in Software Patentability

The patents data are organized and presented in groups spanning multiple years: a 11-year group for the period 1980 to 1990, and groupings of five years after 1990, with figures for 2011 separately reported. We chose 1980 as the starting point in order to trace developments in software patenting after the pivotal US Supreme Court ruling in *Diamond v. Diehr* in 1981 provided the first instance of the USPTO being ordered to grant a patent on an invention which utilized computer software. Prior to this, the USPTO had been reluctant grants patents on inventions relating to computer software, going so far as to issue formal guidelines in 1968 stating that a computer program was non-patentable, whether it was claimed as an apparatus or as a process. The 1981 decision in *Diamond v. Diehr* led the USPTO to modify its position and sparked a period of increased growth in software patenting (Hall and MacGarvie, 2009). Nonetheless, the scope of permissible claims was still relatively limited, requiring software to be contained within a patentable invention. Indeed, it may be argued that "software patents" in this period are more accurately described as "software-related", as they were granted for inventions incorporating software components rather than software inventions per se.

The five-year groupings after 1990 parallel periods of important judicial developments that expanded the scope of computer software as statutory subject matter. In 1991-1995, the Court of Appeals of the Federal Circuit handed down a series of decisions, including *In re Alappat* (1994), *In re Lowry* (1994) and *In re Beauregard* (1995), which determined that much of software was patentable. In response, the USPTO proposed new guidelines in 1995, which were published in

early 1996, stating that it would allow software embedded in physical media to be claimed as processes. In addition to these guidelines, the period 1996-2000 saw further expansion in the scope of permissible software claims when the Federal Circuit's ruling in *State Street Bank & Trust v. Signature Financial Group* (1998) established computer-implemented business methods as patentable.

In the most recent period 2006-2010, a series of decisions have examined the scope of patentable subject matter pertaining to processes, as well as the means of testing the patentability of process claims. These decisions are of particular relevance to software inventions, particularly those that are not coupled or combined with hardware or machinery. Ruling on *In re Bilski* (2008), the Federal Circuit set forth the machine-or-transformation test for process claims, which appears to curtail the patentability of certain software and business methods. However, this decision was partially reversed by the Supreme Court in *Bilski v Kappos* (2010), in which the court rejected the test as the sole test of patentability of processes. In the same ruling, the Supreme Court also rejected the categorical exclusion of business methods from patent eligibility. While this decision allayed fears that *In re Bilski* would signal the end of software process patents, the Supreme Court also emphasized that examiners have the ability to reject method claims as pre-empting an abstract idea. B*ilski v Kappos* is regarded as a landmark decision for software patentability and its ramifications are as yet unclear. As our dataset ends in 2011, we can at best partially assess the impact of this decision on aggregate software patenting trends.

Defining and Identifying Software Patents

In the literature and at patent offices, there are multiple definitions of what constitutes a "software patent." The USPTO's US Patent Classification (USPC) system has devoted a section of classes to “computer implemented patents”, spanning USPC 700 to USPC 726. This is a broad categorization which was further expanded upon by Bessen (2011) to include classes of technologies that are "reliant on software". Among "Computer Implemented Patents" are "Business Methods Patents" which are assigned USPC 705. In the academic literature, scholars have attempted to construct datasets of software patents through keyword searching (Bessen and Hunt, 2007) and identifying patents of top software or ICT firms (Graham and Mowery, 2003; Arora et al., 2007; Hall and MacGarvie, 2010). In a report on the software industry (Lippoldt and Stryszowski, 2009), the OECD utilized the methodology developed by Arora et al. (2007).

Table 2 summarizes the approaches and definitions that have been used in the literature and in practice. Depending on the definition, the number of identified software patents granted by the USPTO in the 32-year period 1980-2011 varies from 178,889 (Bessen and Hunt’s (2007) keyword search) to over 480,000 patents (Bessen’s (2011) expansion of the USPTO’s classification). In this paper, we adopt the two methodologies which are based on technology classes assigned in

the patent document; namely the USPTO's classes USPC 700-726 and the schematic developed by Arora et al. (2007) which uses 5 IPC classes and has been adopted by the OECD.

Table 2 Definitions of Software Patents

	Identification Method	Total USPTO Patents 1980-2011
USPTO	"Computer Implemented Patents" USPC 700-726	399,531
USPTO	"Business Methods Patents" USPC 705	30,315
Graham and Mowery (2003)	3 IPC Classes based on patents of selected software firms: G06F (Electrical Digital Data Processing) G06K (Recognition of Data), H04L (Secure Transmission of Digital Info)	249,818
Bessen and Hunt (2007)	Keyword search ("software" or "computer program", with exclusion words)	178,889
Arora et al. (2007), used by OECD (Lippoldt & Stryskowski, 2009)	5 IPC Classes: 3 as in Graham & Mowery + 2 others G06F (Electrical Digital Data Processing) G06K (Recognition of Data), H04L (Secure Transmission of Digital Info) G06T (Image Data Processing) G09G (Visual Indicators)	254,387
Bessen (2011)	USPC 700-707, 715-717 (data processing) + other selected classes ("reliant on software" and in which software firms patent)	481,288
Hall and MacGarvie (2010)	USPC subclasses based on patents on top ICT firms (details not revealed)	NA

Note: Number of USPTO granted 1980-2011 calculated by authors based on provided definitions

Trends in Global Software Patenting

Regardless of the definition used, the volume of software patenting has increased significantly since the 1980s, with growth accelerating in the last decade, as shown in **Table 3**. In 1980, before the *Diamond v Diehr* decision of 1981, there were relatively few patents with software elements. By 1991, the number of software patents granted annually had increased substantially and continued to grow rapidly over the last 20 years. The number of software patents granted in 2011 is 43,604 when the USPTO definition is used, almost triple the figure of 15,499 granted in 2001. When the Aroral et al. definition is used, the number of software patents granted in 2011 was 28,352, also almost tripling the figure of

9,578 patents granted ten years earlier. In the five year period, 2006 to 2011, software patenting grew at rates ranging from 9% to 11% per annum, depending on the definition adopted. This outpaces the growth in computer software spending which was earlier reported in **Table 1** to average 5.4% over the same five year period, and attests to the dramatic surge in software patenting in the last decade.

Corresponding to this rapid growth, software patents contribute an increasingly large share of patents granted by the USPTO. This is illustrated in **Figure 2**. Using the broadest definition (Bessen, 2011), software patents form 18% of all patents in 2006-10, increasing from a share of 12% in the previous five year period. Using the USPTO's "computer-implemented" definition, software patents constitute 15% of all patents in 2006-2010, increasing from a share of 10% in the period 2001 to 2005.

Table 3 Number and Growth of USPTO-Granted software patents by different definitions of "software"

Year of Grant	Bessen (2011)	USPTO (USPC 7XX)	Arora et al (2007)	Graham & Mowery (2003)	Bessen & Hunt (2007)
1980	1,804	1,247	892	892	232
1991	5,347	3,990	2,811	2,811	1,381
1996	9,739	8,323	5,091	4,974	3,255
2001	18,721	15,499	9,578	9,240	7,555
2006	33,527	27,702	17,272	16,878	12,724
2011	51,962	43,604	28,705	28,352	2,1128
Average annual growth 2006-11	9.2	9.5	10.7	10.9	10.7

Figure 3 expands on **Figure 2** by showing the detailed trend in annual grants of software patents as a share of all patents issued by the USPTO. We observe that in the 1990s and early to mid-2000s, there was generally faster growth in periods following major judicial decisions on software patentability. This is especially apparent when the USPTO definition of software patents is used. Interestingly, there was no easing of growth in the period immediately after the *In re Bilski* (2008) ruling that process claims had to pass the machine or transformation test in order to be patentable. While this ruling on the surface suggested that many software processes and business methods claims would be invalidated, patent attorneys learned to draft software claims as machines or manufactures, rather than as processes. Growth in software patents did appear to ease off slightly in 2011, after the *Bilski v Kappos* decision was handed down in 2010. The *Bilski* decision affirmed that methods and processes may qualify for patent protection, but rejected method claims that are attempts to patent an abstract idea. The

slowing growth of software patents in 2011 suggests that there is uncertainty over how to distinguish between patentable software and business methods, and non-patentable abstract ideas. We are mindful that it is premature at this stage to draw any conclusions or venture a prediction about the future growth of software patents based on a single data point.

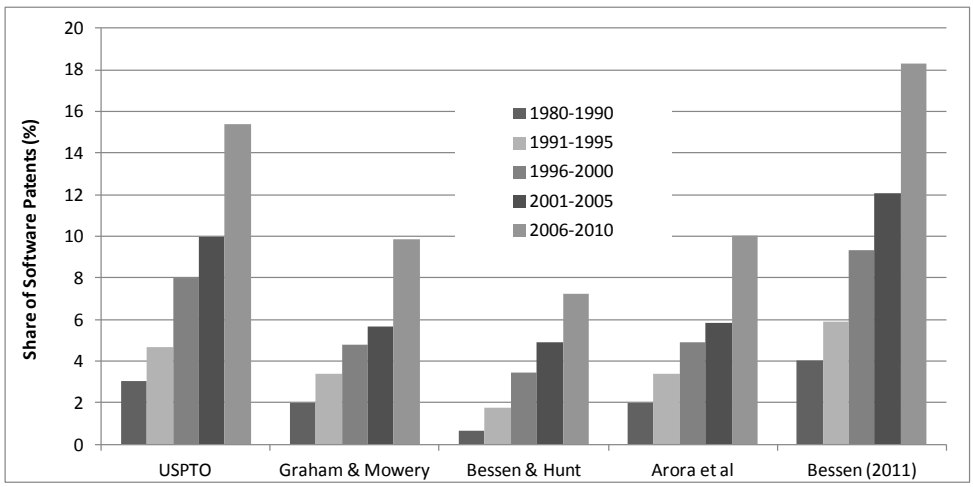


Figure 2 Share of software patents in total USPTO patents by different definitions of "software"

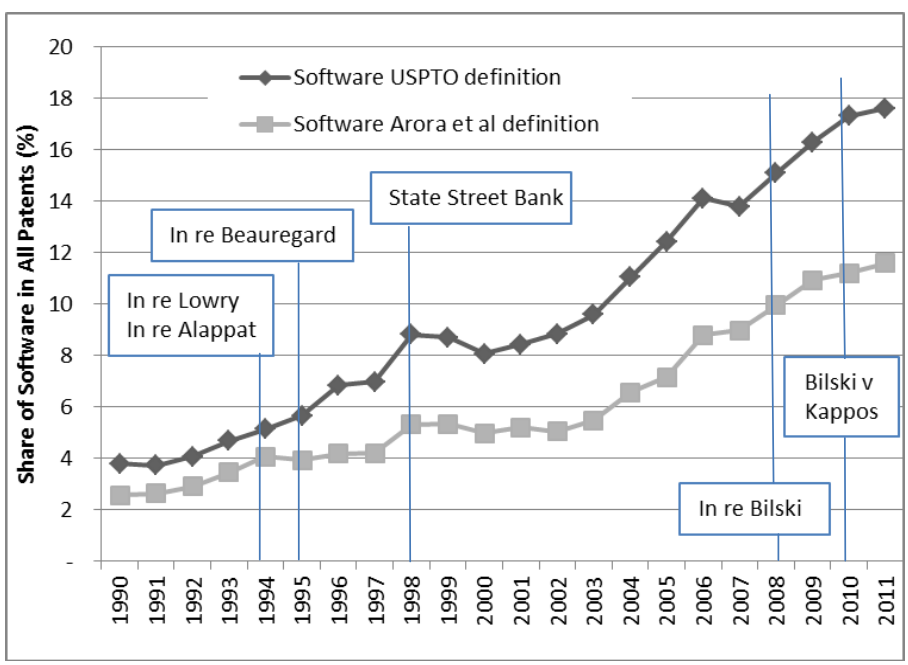


Figure 3 Share of software patents in total USPTO patents, 1990-2011

Table 4 Software Patents Invented in Asia vs Major Software Producing Economies, 1980-2011

	Software USPTO (USPC 7XX) Definition			Software Arora et al. Definition			*ALL PATENTS GRANTED BY USPTO (All Technologies)*		
	Asia with Japan	**Non Japan Asia**	**Major SW Prod.**	**Asia with Japan**	**Non Japan Asia**	**Major SW Prod.**	***Asia with Japan***	***Non Japan Asia***	***Major SW Prod.***
Number of Software Patents invented in Grouping							*Total patents invented in Grouping*		
1980-90	7,213	57	16,926	4,563	31	11,776	*155,676*	*6,031*	*582,677*
1991-95	8,694	396	15,228	6,051	259	11,997	*130,132*	*14,082*	*331,736*
1996-00	14,559	1,823	38,507	9,202	1,079	24,733	*185,822*	*38,268*	*428,421*
2001-05	19,459	3,541	60,005	12,034	1,896	32,526	*245,618*	*66,172*	*493,644*
2006-10	34,799	10,491	107,530	25,157	6,916	68,487	*309,858*	*110,575*	*522,938*
2011	10,024	3,943	32,796	7,997	2,893	20,263	*80,296*	*31,477*	*140,759*
TOTAL	**94,748**	**20,251**	**270,992**	**65,004**	**13,074**	**169,782**	***1.107 m***	***0.267 m***	***2.5 m***
Share of Grouping in Total Software Patents Granted by USPTO (%)							*Share of Grouping in All Patents Granted by USPTO (%)*		
1980-90	27.61	0.22	64.78	26.66	0.18	68.81	*17.98*	*0.70*	*67.3*
1991-95	33.66	1.53	58.96	32.11	1.37	63.67	*23.58*	*2.55*	*60.1*
1996-00	24.16	3.03	63.90	24.96	2.93	67.09	*24.63*	*5.07*	*56.8*
2001-05	21.78	3.96	67.16	22.98	3.62	62.11	*27.46*	*7.40*	*55.2*
2006-10	22.54	6.80	69.66	25.03	6.88	68.15	*30.96*	*11.05*	*52.2*
2011	22.99	9.04	75.21	27.86	10.07	70.59	*32.41*	*12.71*	*56.8*

Note 1: Major Software Producers are North America, Germany, France and United Kingdom
Note 2: Figures refer to patents granted by USPTO in stated period, with at least one inventor from the stated region

Software Patenting in Asia

Region-wide trends

The first part of **Table 4** shows the number of software patents invented in Asia in the last 32 years.[162] As a comparison, **Table 4** also reports figures for a grouping of major software producing economies: North America (USA and Canada), Germany, France and United Kingdom. Using the USPTO definition of software patents, close to 95,000 software patents have been granted to inventors from Asia, the majority of which were granted to Japanese inventors. This compares against nearly 271,000 patents invented in the major software producers grouping. When the more restrictive Arora et al. definition is used, 65,000 software patents have been granted to inventors from Asia. Strikingly, the bulk of Asia's software patents were granted in the last ten years. In the Asian economies other than

[162] "Asia" refers collectively to Japan, China, India, South Korea, Taiwan, Singapore and Hong Kong. The South East Asian economies (eg. Thailand, Philippines) and South Asian economies (eg. Bangladesh, Sri Lanka) are excluded as very few USPTO patents are granted to inventors from these economies.

Japan, more than two thirds of software patents were granted very recently, in the period spanning 2006 to 2011.

The respective contributions of Asia and the major software producers to the world's pool of software patents are reported in the second part of **Table 4**. The group of major software producers have consistently accounted for around two thirds of all software patents granted by the USPTO. The share of this grouping has risen gradually to almost 70% in 2006-2010 and climbed to 75% in 2011 (using the USPTO definition). In contrast, the share of Asian economies exhibits a more dynamic trend. In the last fifteen years up to 2010, around one fifth to one quarter of software patents granted by the USPTO were invented in Asian economies, with Japan contributing much of this share. As shown in **Table 4**, the contribution of Asian economies to software patenting peaked in the period 1991-1995 and has since decreased. This lower share is mainly due to the declining prominence of Japan as a creator of patented software inventions. The share of non-Japan Asia in software patenting has in fact increased quite substantially in the last 20 years, rising from less than 2% in 1991-95 to almost 7% in the period 2006-2010 and over 10% in 2011. However, the contribution of non-Japan Asia to global software patenting is still relatively small.

We recall from **Table 2** that Asia's share in global software spending was 12.9% in 2011 (8.8% excluding Japan). Comparing these figures to the last row in **Table 4**, we observe that non-Japan Asia's contribution to global software patenting is comparable to its share in global software spending. However, Japan's contribution to software patenting is disproportionately higher than its level of software spending.

As a further comparison, the last two columns of **Table 4** report the contribution of each grouping to the overall stock of patents granted by the USPTO across all technology classes. In contrast to the software subset, Asia's share in overall patents has increased steadily over the years. In 2006-2010, Asia-invented patents accounted for 31% of all USPTO patents, while Asia's share of software patents was only 22.5% (USPTO definition) or 25% (Arora et al definition). A similar trend is observed when Japan is excluded. The non-Japan Asian economies contributed 11% of all USPTO patents in 2006-10, but their contribution to the pool of software patents was a more modest 6.9%. However, the contribution of non-Japan Asia to total software patents rose to 10% in 2011. While still lower than the region's share in all patents (12.7%), this shows the continuing expansion of software patenting in non-Japan Asia. In comparison to Asia, figures for the major software producing economies reveal a directly opposite trend. The share of this grouping in all patents has declined steadily, from 67% in the 1980s to 57% in 2011. However, this erosion in share has not taken place in the software sector, where the grouping has strengthened its competitiveness.

Table 5 Average Annual Growth in Software Patents Invented in Asia vs Major Software Producing Economies

	1980-1990	1991-1995	1996-2000	2001-2005	2006-2011
Software USPTO (USPC 7XX) Definition					
Asia with Japan	23.68	12.52	8.82	6.53	15.39
Non Japan Asia	-	68.18	27.8	13.72	27.13
Japan	23.48	10.87	6.68	4.94	10.75
Major SW Prod.	10.28	15.19	20.11	7.16	15.12
World	13.04	14.49	17.05	6.68	14.26
Software Arora et al. Definition					
Asia with Japan	21.32	14.35	7.56	-1.71	27.77
Non Japan Asia	-	62.66	28.73	6.92	38.21
Japan	21.51	12.18	5.44	4.30	15.93
Major SW Prod.	10.21	16.07	16.69	-7.2	29.00
World	12.37	15.12	14.45	5.18	16.79
ALL USPTO PATENTS (All Technologies)					
Asia with Japan	*12.7*	*4.96*	*10.31*	*0.51*	*10.01*
Non Japan Asia	*27.78*	*28.29*	*22.56*	*3.96*	*15.69*
Japan	*12.13*	*2.65*	*7.61*	*-0.69*	*7.26*
Major SW Prod.	*2.88*	*2.32*	*8.8*	*-2.58*	*7.25*
World	*4.60*	*3.52*	*9.09*	*-2.43*	*8.04*

Note 1: Growth rates (%) are for patents granted by USPTO in stated period, with at least one inventor from the stated region

Note 2: Major Software Producers are North America, Germany, France and United Kingdom

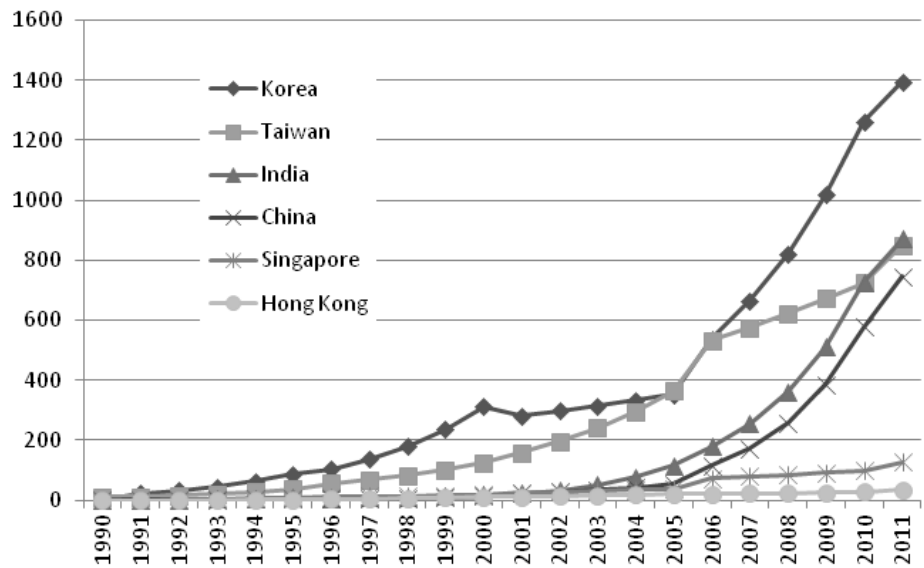
Table 5 reports the growth of software patents invented in Asian economies compared with the grouping of major software producers, using both the USPTO and Arora et al. definitions of software patents. In the early 1990s, very high growth rates were recorded because Asian economies other than Japan were starting their patenting activities from small bases. Prior to 2005, software patenting in Asia as a whole tended to lag behind the global average growth rate. In the period 2006-2011, growth in Asian software patenting outstripped the rest of the world and was almost at par with the group of major software producers. In particular, non-Japan Asia has achieved extremely high growth in the last five year (38% per annum using the Arora et al. definition), higher even the major software producers (29%) that had rebounded very strongly from a downturn in the previous period.

Significantly, **Table 5** also shows that software patents are being invented and granted at a faster rate than patents in other technologies, both globally and specifically in Asia and the major software producing countries. The disparity between software patents growth and growth in other technologies is most pronounced when software patents are identified using the Arora et al. definition. Between 2006 and 2011, Asia's total patent portfolio grew at 10% annually.

Comparably, Asia's portfolio of software patents increased at a much higher rate of 27.8% annually. Similarly, software patents invented in non-Japan Asia grew 38.2% annually in this period, compared to 15.7% growth across all technologies.

Trends in Individual Asian Economies

The rapid growth in software patenting by non-Japan Asia is driven by a few key economies, as shown in **Figure 4**. For convenience, only trends in software patents as defined by USPTO's class 7XX are shown. It is noted that the trends using the Arora et al. definition are very similar and the discussion herein will be equally applicable in that situation.



Note: Figures refer to patents granted by USPTO in stated period, with at least one inventor from the stated economy

Figure 4 Software Patenting in Selected Asian Economies, 1990-2011(USPTO Definition)

As seen **Figure 4**, Korea and Taiwan are the two largest producers of software patents in non-Japan Asia, experiencing a great acceleration that began in the early 2000s and which is still being maintained, albeit at a slower rate in the case of Taiwan. In the mid-2000s, software patenting took off in India and China, with the number of patents granted annually rising very rapidly. By 2010, India has caught up to Taiwan. In 2010, Indian inventors were granted over 700 patents, the figure rising to over 800 in 2011, matching the numbers granted to Taiwanese inventors. China also appears to be quickly catching up to Taiwan and may reach Taiwan's software patents numbers within the next few years. Comparably, the number of software patents produced in Singapore and Hong Kong is relatively small.

As expected, Taiwan and Korea own the largest software patent portfolios among the Asian economies excluding Japan. As seen in **Table 6**, Korean inventors have produced 8,113 software patents using the USPTO definition (4,744 using the Arora et al. definition) while Taiwanese inventors have been granted 5,776 software patents (3,894 using the Arora et al. definition). Comparably, Singapore and Hong Kong have only produced 809 (499) and 265 (137) software patents respectively.

In most of the Asian economies, the majority of software patents in the national portfolio were granted in over the last five to six years. This is especially the case for China and India, where recently granted patents form over 80% of the countries' software patents stock.

Table 6 Number of Software Patents Invented in Asian Economies, 1980-2011

	China	India	Taiwan	Korea	Hong Kong	Singapore	Japan
USPTO (USPC 7XX) Definition							
Up to 1990	5	4	33	5	8	2	7,157
1991-1995	12	18	116	224	9	17	8,302
1996-2000	32	46	519	1,116	39	79	12,741
2001-2005	174	328	1,328	1,493	75	199	15,927
2006-2010	1,520	1,869	2,929	3,879	99	386	24,375
2011	745	871	851	1396	35	126	6,118
TOTAL	2,488	3,136	5,776	8,113	265	809	74,620
Arora et al. Definition							
Up to 1990	6	1	21	1	2	-	4,532
1991-1995	10	11	91	126	9	12	5,795
1996-2000	19	33	333	636	19	45	8,116
2001-2005	103	143	821	583	26	87	10,111
2006-2010	1,162	1,327	1,934	2,401	64	261	18,219
2011	571	605	694	997	17	94	5,108
TOTAL	1,871	2,120	3,894	4,744	137	499	51,881

Note: Figures refer to patents granted by USPTO in stated period, with at least one inventor from the stated economy

Growth in software patenting varies across the different Asian economies. The pattern of growth also varies slightly depending on the definition of software patents adopted, as seen in **Table 7**. For the purposes of discussion, we will focus on the USPTO definition which is more inclusive and covers a larger number of software patents invented in Asia.

After a decade of growth averaging in excess of 20% per annum, Taiwanese software patents eased to a slower rate of 15% growth in the last six years. On the other hand, Korea experienced slowing growth in the early 2000s, when the number of software patents granted increased at a modest 2.5% per annum.

However, Korea bounced back in 2006-2011 to record growth of 26% per annum, outstripping Taiwan.

In the two small NIEs of Singapore and Hong Kong, software patenting has grown at comparably slower rates than the other Asia economies. Hong Kong in particular has not kept pace with the rest of Asia, with annual growth slowing to 9.8% in the last six years.

In the last eleven years, software patenting in China and India grew faster than all the other Asian economies, reflecting the expansion of ICT industries in these two emerging giants. China's growth in the period post 2005 has been especially strong, almost doubling the growth rate achieved in the early to mid-2000s.

Table 7 Average Annual Growth in Software Patents in Asian Economies

	China	India	Taiwan	Korea	Hong Kong	Singapore	Japan
USPTO (USPC 7XX) Definition							
1991-1995	31.61	41.42	31.61	204.53	-	56.51	10.87
1996-2000	16.72	16.27	28.27	29.40	40.63	24.57	6.68
2001-2005	32.45	46.83	23.90	2.50	12.70	15.50	4.94
2006-2011	55.35	39.94	15.15	25.75	9.78	22.66	10.75
Arora et al. Definition							
1991-1995	7.46	-	42.67	91.68	-	-	12.18
1996-2000	4.56	14.87	25.70	36.51	8.45	29.67	5.44
2001-2005	42.13	40.63	15.00	0.77	18.47	17.84	4.30
2006-2011	64.33	49.13	24.88	39.90	15.94	24.70	15.93

Note: Growth rates are for patents granted by USPTO in stated period, with at least one inventor from the stated region

The experiences of these Asian economies contrast with Japan's, as seen in the last column of **Table 7**. While most of the non-Japan Asian economies recorded rapid expansion from the 1990s onwards, Japan's growth rate slowed in the mid-1990s to mid-2000s. From 2006 to 2011, Japan's software patenting picked up pace but still grew at a much slower rate than the other Asian economies excepting Hong Kong. The growth trend in Japan reflects the changing approach to managing software IPR among Japanese companies. In the past, Japanese companies did not attach priority to software and bundled software free with machines and hardware. As a result, the Japanese software industry is underdeveloped compared to other technology sectors. To reverse this, and to capitalize on the opportunities to grow market share in light of the global gravitation towards software patenting, Japanese companies began filing software-related patents in large numbers.

The growth of software patenting has increased the share of software inventions in the national patent stocks of the Asian economies. Again, the discussion will focus on software patents as defined by the USPTO's "computer implemented" categories. It is noted that the trends are similar when the Arora et al. definition is used instead.

Table 8 Share of Software Patents in Total Patents of Asian Economies

	China	India	Taiwan	Korea	Hong Kong	S'pore	Japan	Major SW Prod.	World
USPTO (USPC 7XX) Definition									
1980-90	2.1	2.0	0.9	0.5	0.8	1.6	3.0	2.9	3.0
1991-95	3.4	8.1	1.5	5.3	0.8	5.8	5.0	4.6	4.7
1996-00	4.2	7.6	2.6	7.8	2.0	8.4	5.5	9.0	8.0
2001-05	4.7	16.4	3.9	6.9	2.7	7.8	5.6	12.1	10.0
2006-10	10.7	37.7	6.9	8.5	3.2	12.2	9.1	20.6	15.4
2011	14.9	48.5	8.1	10.4	6.7	14.1	12.5	23.3	17.6
Arora et al. Definition									
1980-90	2.6	0.5	0.6	0.1	0.2	0.0	4.8	2.0	2.0
1991-95	2.8	4.9	1.2	3.0	0.8	4.1	7.1	3.6	3.4
1996-00	2.5	5.5	1.7	4.4	1.0	4.8	8.6	5.8	4.9
2001-05	2.8	7.2	2.4	2.7	0.9	3.4	8.9	6.6	5.9
2006-10	8.2	26.7	4.5	5.3	2.1	8.3	12.2	13.1	10.0
2011	11.4	33.7	6.6	7.4	3.2	10.5	10.4	14.4	11.6

Note 1: For patents granted by USPTO in stated period, with at least one inventor from the stated region

Note 2: Major Software Producers are North America, Germany, France and United Kingdom

A similar trend in observed in China and Singapore, where software patents account for over 10% of recently patented inventions in these two economies, increasing from 4.7% and 7.8% in the previous period 2001-05. While higher than the shares of software patents in Japan, Korea and Taiwan, the shares in China and Singapore are still below the global average.

The exception to this trend among Asian economies is India. Since 2001, the share of software patents in total patents granted to Indian inventors has been higher than the global average. In the five year period 2006-2010, software patents have grown to form a substantial 37.7% of Indian-invented patents, more than double the global average of 15.4%. More recent figures indicate that the share of software patents in India continues to grow rapidly. Examining patents granted to Indian inventors in 2011, nearly half (48.5%) are software-related, almost triple the global average of 17.6%. This affirms the dominance of ICT in India's innovation landscape and the significance of software in India's IP creation activities.

As shown in **Table 8**, software patents account for increasingly large shares of patents granted to inventors from the seven Asian economies. In the two large NIEs of Korea and Taiwan, software patents contribute 8.5% and 6.9% respectively of patents granted in 2006-2010, and 10.4% and 8.1% of patents granted in 2011. These figures have increased from 6.9% and 3.9% in the previous five year period. Nonetheless, the share of software patents is still lower than the global average of 15.4% for 2006-2010 (17.6% for 2011), and much lower than the 20.6% (23.3%) share in the grouping of major software producing countries. We observe a similar trend in Japan, where the share of software patents in the national patent stock has risen substantially from 5.6% in 2001-2005 to 12.5% in 2011, but remains much lower than the global average. While Japanese, Korean and Taiwanese inventors are producing software patents in large numbers, and software patents are increasing at high growth rates, the contribution of software to overall national patenting is still relatively low and has room to grow.

Conclusion

The analysis in this paper highlights the rapid growth of global software patenting over the last ten years, and confirms that this global trend of increasing software patenting is also taking place in Asia. In fact, software patents are growing at a much faster rate in Asian economies than elsewhere. In non-Japan Asia, the growth in software patents even outpaces growth in the major software producing countries. However, excluding Japan, Asian economies still contribute a relatively low share of the world's software patent stock, and with the exception of India, software patents still form a very small proportion of national patents portfolios. In this regard, the Asian economies still lag behind the advanced economies in software patenting, albeit catching up fast.

References

Arora, A., Forman, C. and Yoon, J. (2007). Software. In Jeffrey T. Macher and David C. Mowery (Eds.), *Innovation in Global Industries: American Firms Competing in a New World* (pp 53-99). Washington DC: The National Academies Press.

Bessen, J. (2011). A generation of software patents. *Boston University School of Law Working Paper*, No. 11-31.

Bessen, J. and Hunt, R.M. (2007). An empirical look at software patents. *Journal of Economics and Management Strategy,* 16(1), 157-89.

Graham, S. J. H. and Mowery, D.C. (2003). Intellectual property protection in the U. S. software industry. In Wesley M. Cohen and Stephen A. Merrill (Eds.), *Patents in the Knowledge-Based Economy* (pp. 219-58). National Research Council, Washington: National Academies Press,.

Hall, B. and MacGarvie, M. (2010). The private value of software patents, *Research Policy*, 39(7), 994-1009.

Jafee, A.B. and Lerner, J. (2006). *Innovation and its Discontents: How our Broken Patent System is Endangering Innovation and Progress, and What to do About it*. Princeton: Princeton University Press.

Lippoldt, D. and Stryszowski, P. (2009). *Innovation in the Software Sector*. Paris: OECD Publishing.

von Graevenitz, G., Wagner, S. & Harhoff, D. (2011). Incidence and Growth of Patent Thickets – The Impact of Technological Opportunities and Complexity. CEPR Discussion Paper number 6900. London: Centre for Economic Policy Research.

SUPPLY AND DEMAND IN SCHOLARLY PUBLISHING: AN ANALYSIS OF FACTORS ASSOCIATED WITH JOURNAL ACCEPTANCE RATES (RIP)

Cassidy R. Sugimoto,[1] Vincent Larivière[2] and Blaise Cronin[1]

[1] *{sugimoto, bcronin}@indiana.edu*
School of Library and Information Science, Indiana University Bloomington (USA)

[2] *vincent.lariviere@umontreal.ca*
École de bibliothéconomie et des sciences de l'information, Université de Montréal; Observatoire des Sciences et des Technologies (OST), Centre Interuniversitaire de Recherche sur la Science et la Technologie (CIRST), Université du Québec à Montréal, Montreal (Canada)

Abstract
There are many indicators of journal quality and prestige. Although acceptance rates are discussed anecdotally, there has been little systematic exploration of the relationship between acceptance rates and other journal characteristics. This study examines such relationships for a set of 1,273 journals across multiple domains. The results suggest that acceptance rate is indirectly correlated with citation-based indicators and directly correlated with journal age. These relationships are most pronounced in the most prestigious journals and vary by discipline.

Conference Topic
Scientometric Indicators (Topic 1), Old and New Data Sources for Scientometric Studies: Coverage, Accuracy and Reliability (Topic 2)

Introduction
The scholarly publication system operates on the basis of exchange. As in any market, there are suppliers (authors) and buyers (journals) of goods (papers). Authors typically want their papers to appear in high profile, high prestige, high impact journals. This signals the value of their goods to the marketplace and, ultimately, increases their stock of symbolic capital (Bourdieu, 1984; Birnhotz, 2006). By publishing high impact, highly cited, highly influential papers, a journal signals *its* worth to the marketplace and thereby reinforces *its* reputation and attractiveness to the marketplace. As a general rule, the best authors want to publish in the best journals, and the best journals want the best authors (those with the potentially best papers) to publish with them. In a perfect (i.e., optimally efficient) market, the best papers would gravitate to the best journals (Oster, 1980). But in this as so many other markets both suppliers and buyers lack perfect information. Absent perfect information, the various actors involved rely upon a

range of indicators (bibliometric, sociometric, demographic) to guide decision-making.

Space at the upper end of the market is highly sought after and in limited supply. Competition and ambition often drive scholars to submit papers to journals beyond their reach, creating a cascade of rejected papers that puts added pressure on reviewers and editors (Cronin & McKenzie, 1992; Kravitz & Baker, 2011; Craig, 2010). Economic models have been proposed to analyze, inter alia, research spillover effects, duality in scientific discovery and congestion in information processing (Besancenot, Huynh, & Vranceanu, 2009, p. 1). Such models highlight the "informational frictions" that occur when papers are being matched with journals (Besancenot, Huynh, & Vranceanu, 2009, p. 2). Peer review is the established mechanism for allocating space to papers. Experts (editors, editorial board members and external reviewers) assess the quality of submitted papers and evaluate their suitability for publication. It is assumed that editors and reviewers are unbiased in their assessments and that the governing norm of impartiality is not violated (Lee, Sugimoto, Zhang & Cronin, 2013). In reality, it is not quite so straightforward, as variations in consensus as to what constitutes quality, broadly conceived, within and across fields, can have an effect on acceptance rates (Hargens, 1988; Kravitz, Franks, Feldman, Gerrity, Byrne & Tierney, 2010).

Variation in journal acceptance rates is an understudied area, not the least because of the difficulty in obtaining reliable data. One of the most comprehensive studies to date examined the rejection rates of 83 journals across a broad spectrum of disciplinary areas and found that humanities and social science journals have the highest rates and the biological sciences the lowest (Zuckerman & Merton, 1971, p. 77): "the more humanistically oriented the journal, the higher the rate of rejecting manuscripts for publication; the more experimentally oriented, with an emphasis on rigour of observation and analysis, the lower the rate of rejection." Subsequent monodisciplinary studies have confirmed these findings (e.g., Cherkashin, Demidova, Imai, & Krishna, 2009; Seaton, 1975; Vlachy, 1981; Rotton, Levitt, & Foos, 1993; Schultz, 2010). One explanation of this is the degree to which a dominant paradigm exists in the given discipline, providing a consensus as to what constitutes valid research (Kuhn, 1970).

The complexity of this market exchange system and the amount of variation in acceptance rates raise issues of reliability and validity. It has been noted that there exists little guidance for computing acceptance rates (Moore & Perry, 2012). At face value, the calculation may seem simple enough—the number of papers accepted over the total number of papers submitted. However, this is complicated by the unreliability of self-report data, the inconsistent definitions of a resubmission, the inclusion/exclusion of invited papers or special issues in the calculations, the timeframe used, and the inclusion/exclusion of book reviews, among other considerations (Moore & Perry, 2012). Additionally, many studies rely on individual surveys of editors/publishers, rather than using a standard source for evaluation. Cabell's Directories of Publishing Opportunities (Cabell's

henceforth) is one such source, but has been used only rarely (e.g., Haensly, Hodges, & Davenport, 2008).

Acceptance rates testify to the relative competitiveness of a journal, but have also been used as a quality measure. Significant indirect correlations between acceptance rates and other proxies of quality (i.e., citation rates, Journal Impact Factor [JIF]) have been demonstrated (Lee, Schotland, Bacchetti, & Bero, 2002; Buffardi & Nichols, 1981; Haensley, Hodges, & Davenport, 2008). However, and with few exceptions, these have relied on small scale and monodisciplinary datasets and are somewhat dated. Rotton, Levitt, and Foos 1993) found that rejection rates were good predictors of citations, while Haensly, Hodges, and Davenport (2008) found acceptance rates to be significantly correlated with both citations and survey-based rankings of journals. However, they also noted that circulation was one of the most important predictors of quality (JIF, rejection rates, etc.). Here we examine acceptance rates for all journals listed in Cabell's. This is the largest study of acceptance rates to date and provides a cross-disciplinary analysis of the relationship between acceptance rates and other journal quality measures.

Table 1. Number of unique journals and percent indexed in the JCR by discipline

Discipline	Specialty	# of unique journals	# of unique journals in JCR	% of journals in JCR
Business	Accounting	464	35	7.5%
	Management	1,652	286	17.3%
	Marketing	217	24	11.1%
	Economics and Finance	1,221	285	23.3%
	Subtotal	***2,628***	***555***	***21.1%***
Computer Science	*Computer Science and Business Information Systems*	***771***	***154***	***20.0%***
Education	Educational Technology and Library Science	295	37	12.5%
	Educational Curriculum and Methods	626	108	17.3%
	Educational Psychology and Administration	521	113	21.7%
	Subtotal	***1,215***	***235***	***19.3%***
Health	Nursing	235	73	31.1%
	Health Administration	236	77	32.6%
	Subtotal	***351***	***118***	***33.6%***
Psychology	Psychology and Psychiatry	***779***	***479***	***61.5%***
TOTAL		**5,092**	**1,348**	**26.5%**

Methods

We used three main sources of data: Cabell's, Thomson Reuters' Journal Citation Reports (JCR), and Ulrich's Periodicals Directory. Cabell's provides general descriptive information about journals (Cabell's Directories, 2012). It indexes

journals in eleven specialties organized into five disciplines (Table 1). Each journal can be assigned to multiple specialties. New journals can be recommended to the directory by emailing a form to the company. Journal information is obtained by contacting editors and/or publishers, but is not independently verified.

Basic metadata for all the journals by specialty was downloaded from Cabell's, including acceptance rate and whether the journal was indexed in JCR. In total, 7,015 records were downloaded for all specialties; these represented 5,092 unique journals, as journals appear in multiple specialties (Table 1). Of these, more than one quarter (N=1,348) of unique journals were listed as indexed in JCR.

The 2011 data for both the Science and Social Sciences Journal Citation Reports were downloaded from Thomson Reuters (variables collected are shown in Table 1). However, all of these data were associated with the abbreviated name of the journal. A conversion table and title matching were used to match the Cabell's and JCR data. The JCR data were located for 1,273 unique journals. Journals could not be matched for several reasons: 1) incomplete information was provided in Cabell's (e.g., the title did not include the subtitle, so could not be distinguished from several journals with the same initial title); 2) no such journal could be found in JCR (due either to an erroneous assumption on the part of the editor or because the journal had ceased to appear in JCR since the time the editor was surveyed); and 3) the journal was indexed in the Humanities index of Web of Science for which Impact Factors are not calculated. In order to account for different citation practices across disciplines, we also compiled field-normalized Impact Factors, which were obtained by dividing the impact factor of each journal by the average impact factor of papers published in the same discipline. Lastly, journal start dates were gathered from Ulrich's Periodicals Directory using the "Start Year" field in the database. In the case of journal name changes, the start date of the initial journal was used, not the date at which the journal became associated with the current name.

Results and Discussion

All 1,273 unique journals found in the JCR were analyzed (two-tailed Pearson correlation analysis) to examine the relationships between journal factors and acceptance rates (Table 2).

As can be seen, there was an indirect relationship between acceptance rates and all citation-related metrics—that is, when the acceptance rate decreases, the citation rate tends to increase. There was a direct relationship between the number of articles and the acceptance rate—that is, acceptance increased as the number of articles in a given journal increased. There was also a direct relationship between start year and acceptance rate—that is, younger journals tended to have higher acceptance rates. The strongest correlations were between acceptance rates and cited half-life and acceptance rates, article influence score and field-normalized IFs. There was no significant relationship between total cites; all other relationships were significant at either the .05 or .01 level.

Table 2. Correlations between journal measures and acceptance rates[163]

	AcceptLow	AcceptMed	AcceptHigh
2011TotalCites (n=1215)	-.039	-.045	-.052
ImpactFactor (n=1213)	-.059*	-0.63*	-.066*
5YearImpactFactor (n=1008)	-.098**	-.106**	-.113**
ImmediacyIndex (n=1212)	-.070*	-.068*	-.065*
2011Articles (n=1212)	.121**	.123**	.123**
CitedHalfLife (n=860)	-.261**	-.258**	-.251**
EigenfactorScore (n=1215)	-.057*	-.064*	-.071*
ArticleInfluenceScore (n=1008)	-.219**	-.230**	-.236**
StartYear (n=1232)	.126**	.131**	.134**
Field normalized Impact Factor (n=1189)	-.193**	-.204**	-.213**

*. Correlation is significant at the 0.05 level (2-tailed)

**. Correlation is significant at the 0.01 level (2-tailed)

Table 3. Correlation between journal factors and median acceptance rates

Variable	1st quartile	2nd quartile	3rd quartile	4th quartile
2011TotalCites	-.175** (n=304)	.015 (n=310)	.021 (n=305)	-.082 (n=296)
ImpactFactor	-.171** (n=304)	-.097 (n=309)	.048 (n=305)	-.012 (n=295)
5YearImpactFactor	-.155 (n=260)	-.114 (n=271)	.075 (n=255)	-.023 (n=222)
ImmediacyIndex	-.085 (n=302)	-.127* (n=310)	-.006 (n=305)	-.006 (n=295)
2011Articles	-.080 (n=185)	.048 (n=310)	.113* (n=305)	.003 (n=295)
CitedHalfLife	-.070 (n=185)	-.223** (n=212)	-.065 (n=229)	-.214** (n=234)
EigenfactorScore	-.209** (n=304)	.012 (n=310)	.066 (n=305)	-.077 (n=296)
ArticleInfluenceScore	-.232** (n=206)	-.185** (n=271)	.027 (n=255)	-.017 (n=222)
StartYear	.153** (n=314)	.046 (n=305)	.111 (n=306)	.196** (n=307)

*. Correlation is significant at the 0.05 level (2-tailed)

**. Correlation is significant at the 0.01 level (2-tailed)

In order to ensure that elite journals were not skewing the results, we divided the journals into quartiles by acceptance rate and analyzed each quartile (Table 3). The first quartile contains those journals with the lowest acceptance rates; the fourth quartile those with the highest. The strongest relationships between the indicators and acceptance rates are in the first quartile, particularly in terms of citation-based indicators, suggesting that there is a relationship between highly competitive journals (those with low acceptance rates) and high impact journals (those with high citation counts). However, the relationship is much more nuanced for less competitive journals: as shown in Table 3, very few significant relationships are found below the first quartile. Interesting exceptions are the

[163]Acceptance rates were given in various forms. While some editors provided an exact percentage (e.g., 17%) others provided a range (e.g., 10-15%). Therefore, prior to analysis, this field was expanded to three: minimum, median, and maximum. For each analysis, a sensitivity analysis was performed, to ensure that different results would not be achieved using one of these three.

significance of number of articles (i.e., 2011Articles) and acceptance rate in the 3rd quartile as well as the cited half-life and acceptance rates in the 4th quartile.
The relationships between journal factors and acceptance rates were also analyzed by discipline (Table 4). Acceptance rates are significantly indirectly correlated with article influence scores across all disciplines. The five-year IF is significantly indirectly correlated with all disciplines, except Psychology, and the cited-half life is significantly indirectly correlated with all disciplines except computer science, where it is a direct relationship—suggesting perhaps that for computer science rigor is associated with speed. The lack of a significant relationship between acceptance rates and start years for computer science and business demonstrates that journal age is not as important in these fields as in others. The JIF is significantly indirectly related to acceptance rates in Business, Computer Science, and Health, but not in Education or Psychology.

Table 4. Relationships between journal factors and acceptance rates by discipline

Variable	Business	Computer Science	Education	Health	Psychology
2011TotalCites	-.145** (n=510)	-.106 (n=131)	-.095 (n=214)	-.316** (n=105)	-.036 (n=433)
ImpactFactor	-.267** (n=510)	-.307** (n=131)	-.132 (n=214)	-.433** (n=104)	-.050 (n=432)
5YearImpactFactor	-.279** (n=408)	-.407** (n=108)	-.159* (n=186)	-.453** (n=78)	-.078 (n=390)
ImmediacyIndex	-.223** (n=509)	-.155 (n=130)	-.046 (n=213)	-.200* (n=105)	-.099* (n-432)
2011Articles	.067 (n=509)	.019 (n=130)	.181** (n=213)	-.070 (n=105)	.164** (n=432)
CitedHalfLife	-.166** (n=330)	.043 (n=103)	-.284** (n=145)	-.252* (n=95)	-.326** (n=302)
EigenfactorScore	-.174** (n=510)	-.109 (n=131)	-.108 (n=214)	-.305** (n=105)	.001 (n=433)
ArticleInfluenceScore	-.287** (n=408)	-.406** (n=108)	-.256** (n=186)	-.485** (n=78)	-.152** (n-390)
StartYear	.080 (n=523)	.010 (n=140)	.161* (n=212)	.213* (n=110)	.181** (n=429)

Conclusions

Most authors would like to see their work appear in a prestigious journal such as *Nature*, but probabilistically that is not going to happen. In all likelihood authors have an intuitive sense of how journals in their field stack up—if they don't there is a growing number of discipline-specific rankings on which to rely (e.g., Harris, 2008)—in terms of reputation and quality and will take a path somewhere between idealism and pragmatism when it comes to submitting their papers, neither aiming too high (waste of time and effort) nor setting their sights too low (bad from a career advancement perspective and for one's morale). In any case,

rejected papers, which will likely have had some value added as a result of their being subjected to peer review, will eventually find a home elsewhere, albeit at a lower level in the overall journal pecking order (e.g., Bornmann, Weymuth, & Daniel, 2010; Sugimoto & Cronin, 2013; Cronin, 2012), but also, sometimes, in higher-end journals (Calcagno et al., 2012). Not everyone can publish in *Nature*; not everyone should try. Unrealistic expectations, when scaled up, translate into system inefficiencies and disequilibria, which is bad news all round.

Space is consistently mentioned in studies of acceptance rates (e.g., Zuckerman & Merton, 1971). In a longitudinal study, Hargens (1988) found that rejection rates remained stable, despite increases in submissions, thereby challenging the space argument. However, our research finds a direct correlation between number of articles published per year and acceptance rates, suggesting that space does in fact play some role. The issue is further complicated by the dramatic changes taking place to the traditional scholarly publishing business model. The introduction of author processing charges is shifting the burden of payment from the consumer (individual or institutional) to the author (Solomon & Björk, 2012). Although open access (OA) has not fundamentally altered the exchange of space for content, it does require us, as Suber (2008) has observed, to think more closely about the relationship between journal *prestige* and journal *quality* in an evolving, mixed-mode publishing market, i.e., one combining toll access and open access journals. As a result of accelerating developments in OA and the rapid adoption of alternative metrics of scholarly influence and impact (e.g., Cronin & Sugimoto, in press) we are seeing an expansion of the indicator set that can be used to assess journal quality. Prestige and Journal Impact Factor are now only two of the indicators available to authors to evaluate journal quality and inform their submission behaviors (e.g., Lozano, Larivière & Gingras, 2012). Work has also been done to explore the relationship between cost and prestige (eigenfactor.org, 2013). By the same token, journals can deploy a wider range of indicators to signal *their* quality to prospective authors and readers. The market may still be imperfect, but it is becoming more transparent.

References

Besancenot, D., Huynh, K., & Vranceanu, R. (2009). Desk rejection in an academic publication market model with matching frictions. DR 09008. ESSEC Business School.

Birnholtz, J. (2006). What does it mean to be an author? The intersection of credit, contribution and collaboration in science. *Journal of the American Society for Information Science and Technology*, *57*(13), 1758–1770.

Bornmann, L., Weymuth, C., & Daniel, H. -D. (2010). A content analysis of referees' comments: How do comments on manuscripts rejected by a high-impact journal and later published in either a low- or high-impact journal differ? *Scientometrics*, *83*, 493-506.

Bourdieu, P. (1984). *Distinction: A social critique of the judgement of taste*. *Cambridge*, MA: Harvard University Press.

Buffardi, L. C., & Nichols, J. A. (1981). Citation impact, acceptance rate, and APA journals. *American Psychologist*, *36*(11), 1455-1456.
Calcagno, V., Demoinet, E., Gollner, K., Guidi, L., Ruths, D., de Mazancourt, C. (2012). Flows of research manuscripts among scientific journals reveal hidden submission patterns. *Science*, *338*(6110), 1065-1069
Cherkashin, I., Demidova, S., Imai, S., & Krishna, K. (2009). The inside scoop: Acceptance and rejection at the Journal of International Economics. *Journal of International Economics*, *77*, 120-132.
Craig, J. B. (2010). Desk rejection: How to avoid being hit by a returning boomerang. *Family Business Review*, *23*(4), 306-309.
Cronin, B., & McKenzie, G. (1992). The trajectory of rejection. *Journal of Documentation*, *48*(3), 310-317.
Cronin, B. & Sugimoto, C. R. (Eds.). (in press). Bibliometrics and beyond: Metrics-based evaluation of scholarly research. Cambridge, MA: MIT Press.
Cronin, B. (2012). Editorial. Do me a favor. *Journal of the American Society for Information Science and Technology*, *63*(7), 1281.
Eigenfactor.org. (2013). Cost effectiveness for open access journals. Retrieved from: http://www.eigenfactor.org/openaccess/
Haensly, P. J., Hodges, P. E., & Davenport, S. A. (2008). Acceptance rates and journal quality: An analysis of journals in economics and finance. *Journal of Business & Finance Librarianship*, *14*(1), 2-31.
Hargens, L. L. (1988). Scholarly consensus and journal rejection rates. *American Sociological Review*, *53*(1), 139-151.
Harris, C. (2008). Ranking the management journals. *Journal of Scholarly Publishing*, *39*(4), 373-409.
Kravitz, D.J., & Baker, C.I. (2011). Toward a new model of scientific publishing: Discussion and a proposal. Frontiers in Computational Neuroscience, 5(55).
Kravitz, R. L., Franks, P., Feldman, M. D., Gerrity, M., Byrne, C., & Tierney, W. M. (2010), Editorial peer reviewers' recommendations at a general medical journal: Are they reliable and do editors care? *PLoS One*, *59*(4), e10072.
Kuhn, T. S. (1970). *The Structure of Scientific Revolutions*. Chicago: University of Chicago Press.
Lee, C. J., Sugimoto, C. R., Zhang, G., & Cronin, B. (2013). Bias in peer review. *Journal of the American Society for Information Science and Technology, 64*(1), 2-17.
Lee, K. P., Schotland, M., Bacchetti, P., & Bero, L. A. (2002). Association of journal quality indicators with methodological quality of clinical research articles. *JAMA*, *287*(21), 2805-2808.
Lozano, G. A., Larivière, V., & Gingras, Y. (2012). The weakening relationship between the impact factor and papers' citations in the digital age. *Journal of the American Society for Information Science & Technology*, *63*(11), 2140-2145.

Moore, M. A., & Perry, S.D. (2012). Oughts v. Ends: Seeking an ethical normative standard for journal acceptance rate calculation methods. *Journal of Academic Ethics*, *10*, 113-121.
Oster, S. (1980). The optimal order for submitting manuscripts. *The American Economic Review*, *70*(3), 444-448.
Rotton, J., Levitt, M., Foos, P. (1993). Citation impact, rejection rates, and journal values. *American Psychologist*, 911-912.
Schultz, D. M. (2010). Rejection rates for journals publishing in the atmospheric sciences. *American Meteorological Society*, 231-243.
Seaton, H. W. (1975). Education journals: Publication lags and acceptance rates. *Educational Research*, *4*(4), 18-19.
Solomon, D. J. & Björk, B. C. (2012). A study of open access journals using article processing charges. *Journal of the American Society for Information Science & Technology*, *63*(8), 1485-1495.
Suber, P. (2008). Thinking about prestige, quality, and open access. SPARC Open Access Newsletter, 125. Available online at: http://www.earlham.edu/~peters/fos/newsletter/09-02-08.htm
Vlachy, J. (1981). Refereeing and rejection patterns in physics journals. *Czechoslovakian Journal of Physics*, B31, 453-456.
Zuckerman, H. & Merton, R. K. (1971). Patterns of evaluation in science: Institutionalisation, structure and functions of the referee system. *Minerva*, *9*(1), 66-100.

A SYSTEMATIC EMPIRICAL COMPARISON OF DIFFERENT APPROACHES FOR NORMALIZING CITATION IMPACT INDICATORS

Ludo Waltman and Nees Jan van Eck

{waltmanlr, ecknjpvan}@cwts.leidenuniv.nl
Centre for Science and Technology Studies, Leiden University, Leiden (The Netherlands)

Abstract
We address the question how citation-based bibliometric indicators can best be normalized to ensure fair comparisons between publications from different scientific fields and different years. In a systematic large-scale empirical analysis, we compare a normalization approach based on a field classification system with three source normalization approaches. We pay special attention to the selection of the publications included in the analysis. Publications in national scientific journals, popular scientific magazines, and trade magazines are not included. Unlike earlier studies, we use algorithmically constructed classification systems to evaluate the different normalization approaches. Our analysis shows that a source normalization approach based on the recently introduced idea of fractional citation counting does not perform well. Two other source normalization approaches generally outperform the classification-system-based normalization approach that we study. Our analysis therefore offers considerable support for the use of source-normalized bibliometric indicators.

Conference Topic
Scientometrics Indicators (Topic 1).

Introduction
Citation-based bibliometric indicators have become a more and more popular tool for research assessment purposes. In practice, there often turns out to be a need to use these indicators not only for comparing researchers, research groups, departments, or journals active in the same scientific field or subfield but also for making comparisons across fields. Performing between-field comparisons is a delicate issue. Each field has its own publication, citation, and authorship practices, making it difficult to ensure the fairness of between-field comparisons. In some fields, researchers tend to publish a lot, often as part of larger collaborative teams. In other fields, collaboration takes place only at relatively small scales, usually involving no more than a few researchers, and the average publication output per researcher is significantly lower. Also, in some fields, publications tend to have long reference lists, with many references to recent work. In other fields, reference lists may be much shorter, or they may point mainly to older work. In the latter fields, publications on average will receive only

a relatively small number of citations, while in the former fields, the average number of citations per publication will be much larger.

In this paper, we address the question how citation-based bibliometric indicators can best be normalized to correct for differences in citation practices between scientific fields. Hence, we aim to find out how citation impact can be measured in a way that allows for the fairest between-field comparisons.

In recent years, a significant amount of attention has been paid to the problem of normalizing citation-based bibliometric indicators. Basically, two streams of research can be distinguished in the literature. One stream of research is concerned with normalization approaches that use a field classification system to correct for differences in citation practices between scientific fields. In these normalization approaches, each publication is assigned to one or more fields and the citation impact of a publication is normalized by comparing it with the field average. Research into classification-system-based normalization approaches started in the late 1980s and the early 1990s. Recent contributions to this line of research were made by, among others, Crespo, Herranz, Li, and Ruiz-Castillo (2012), Crespo, Li, and Ruiz-Castillo (2012), Radicchi and Castellano (2012c), Radicchi, Fortunato, and Castellano (2008), and Van Eck, Waltman, Van Raan, Klautz, and Peul (2012).

The second stream of research studies normalization approaches that correct for differences in citation practices between fields based on the referencing behavior of citing publications or citing journals. These normalization approaches do not use a field classification system. The second stream of research was initiated by Zitt and Small (2008), who introduced the audience factor, an interesting new indicator of the citation impact of scientific journals. Other contributions to this stream of research were made by Glänzel, Schubert, Thijs, and Debackere (2011), Leydesdorff and Bornmann (2011), Leydesdorff and Opthof (2010), Leydesdorff, Zhou, and Bornmann (2013), Moed (2010), Waltman and Van Eck (in press), Waltman, Van Eck, Van Leeuwen, and Visser (2013), Zhou and Leydesdorff (2011), and Zitt (2010, 2011). Zitt and Small referred to their proposed normalization approach as 'fractional citation weighting' or 'citing-side normalization'. Alternative labels introduced by other authors include 'source normalization' (Moed, 2010), 'fractional counting of citations' (Leydesdorff & Opthof, 2010), and 'a priori normalization' (Glänzel et al., 2011). Following our earlier work (Waltman & Van Eck, in press; Waltman et al., 2013), we will use the term 'source normalization' in this paper.

Which normalization approach performs best is still an open issue. Systematic large-scale empirical comparisons of normalization approaches are scarce, and as we will see, such comparisons involve significant methodological challenges. Studies in which normalization approaches based on a field classification system are compared with source normalization approaches have been reported by Leydesdorff, Radicchi, Bornmann, Castellano, and De Nooy (in press) and Radicchi and Castellano (2012a). In these studies, classification-system-based normalization approaches were found to be more accurate than source

normalization approaches. However, as we will point out later on, these studies have important methodological limitations. In an earlier paper, we have compared a classification-system-based normalization approach with a number of source normalization approaches (Waltman & Van Eck, in press). The comparison was performed in the context of assessing the citation impact of scientific journals, and the results seemed to be in favor of some of the source normalization approaches. However, because of the somewhat non-systematic character of the comparison, the results must be considered of a tentative nature.

Building on our earlier work (Waltman & Van Eck, in press), we present in this paper a systematic large-scale empirical comparison of normalization approaches. The comparison involves one normalization approach based on a field classification system and three source normalization approaches. In the classification-system-based normalization approach, publications are classified into fields based on the journal subject categories in the Web of Science bibliographic database. The source normalization approaches that we consider are based on the audience factor approach of Zitt and Small (2008), the fractional citation counting approach of Leydesdorff and Opthof (2010), and our own revised SNIP approach (Waltman et al., 2013).

Our methodology for comparing normalization approaches has three important features not present in earlier work by other authors. First, rather than simply including all publications available in a bibliographic database in a given time period, we exclude as much as possible publications that could distort the analysis, such as publications in national scientific journals, popular scientific magazines, and trade magazines. Second, in the evaluation of the classification-system-based normalization approach, we use field classification systems that are different from the classification system used by the normalization approach itself. In this way, we ensure that our results do not suffer from a bias that favors classification-system-based normalization approaches over source normalization approaches. Third, we compare normalization approaches at different levels of granularity, for instance both at the level of broad scientific disciplines and at the level of smaller scientific subfields. As we will see, some normalization approaches perform well at one level but not so well at another level.

The organization of this paper is as follows. We first discuss the data that we use in our analysis, and we then introduce the normalization approaches that we study. Next, we present the results of our analysis, and finally, we summarize our conclusions. We note that a more extensive version of this paper is available online (Waltman & Van Eck, 2013).

Data

Our analysis is based on data from the Web of Science (WoS) bibliographic database. We use the Science Citation Index Expanded, the Social Sciences Citation Index, and the Arts & Humanities Citation Index. The data that we work with is from the period 2003–2011.

The WoS database is continuously expanding. Nowadays, the database contains a significant number of special types of sources, such as scientific journals with a strong national or regional orientation, trade magazines (e.g., *Genetic Engineering & Biotechnology News*, *Naval Architect*, and *Professional Engineering*), business magazines (e.g., *Forbes* and *Fortune*), and popular scientific magazines (e.g., *American Scientist*, *New Scientist*, and *Scientific American*). As we have argued in an earlier paper (Waltman & Van Eck, 2012), a normalization for differences in citation practices between scientific fields may be distorted by the presence of these special types of sources in one's database. For this reason, we do not simply include all WoS-indexed publications in our analysis. Instead, we include only publications from selected sources, which we refer to as WoS core journals. In this way, we intend to restrict our analysis to the international scientific literature covered by the WoS database. The details of our procedure for selecting publications in WoS core journals are discussed in Appendix A in the more extensive version of this paper (Waltman & Van Eck, 2013). Of the 9.79 million WoS-indexed publications of the document types *article* and *review* in the period 2003–2011, there are 8.20 million that are included in our analysis.

In the rest of this paper, the term 'publication' always refers to our selected publications in WoS core journals. Also, when we use the term 'citation' or 'reference', both the citing and the cited publication are assumed to belong to our set of selected publications in WoS core journals. Hence, citations originating from non-selected publications or references pointing to non-selected publications play no role in our analysis.

Table 1. Summary statistics for each of the four field classification systems.

	No. of areas	Number of publications per area (2007–2010)			
		Mean	Median	Minimum	Maximum
WoS subject categories	235	27,524	16,448	94	191,790
Classification system A	21	182,133	137,548	49,577	635,209
Classification system B	161	23,757	19,085	4,800	69,816
Classification system C	1,334	2,867	2,421	820	12,037

The analysis that we perform focuses on calculating the citation impact of publications from the period 2007–2010. There are 3.86 million publications in this period. For each publication, citations are counted until the end of 2011.

We use four different field classification systems in our analysis. One is the well-known system based on the WoS journal subject categories. In this system, a publication can belong to multiple research areas. The other three classification systems have been constructed algorithmically based on citation relations between publications. These classification systems, referred to as classification systems A,

B, and C, differ from each other in their level of granularity. Classification system A is the least detailed system and consists of only 21 research areas. Classification system C, which includes 1,334 research areas, is the most detailed system. In classification systems A, B, and C, a publication can belong to only one research area. We refer to Appendix B in the more extensive version of this paper (Waltman & Van Eck, 2013) for a discussion of the methodology that we have used for constructing classification systems A, B, and C. The methodology is largely based on an earlier paper (Waltman & Van Eck, 2012).
Table 1 provides some summary statistics for each of our four field classification systems. These statistics relate to the period 2007–2010.

Normalization approaches

As already mentioned, we study four normalization approaches in this paper, one based on a field classification system and three based on the idea of source normalization. In addition to correcting for differences in citation practices between scientific fields, we also want our normalization approaches to correct for the age of a publication. Recall that our focus is on calculating the citation impact of publications from the period 2007–2010 based on citations counted until the end of 2011. This means that an older publication, for instance from 2007, has a longer citation window than a more recent publication, for instance from 2010. To be able to make fair comparisons between publications from different years, we therefore need a correction for the age of a publication.
We start by introducing our classification-system-based normalization approach. In this approach, we calculate for each publication a *normalized citation score* (NCS). The NCS value of a publication is given by

$$\mathrm{NCS} = \frac{c}{e} \tag{1}$$

where c denotes the number of citations of the publication and e denotes the average number of citations of all publications in the same field and in the same year. Interpreting e as a publication's expected number of citations, the NCS value of a publication is simply given by the ratio of the actual and the expected number of citations of the publication. An NCS value above (below) one indicates that the number of citations of a publication is above (below) what would be expected based on the field and the year in which the publication appeared.
To determine a publication's expected number of citations e in (1), we need a field classification system. In practical applications of the classification-system-based normalization approach, the journal subject categories in the WoS database are often used for this purpose. We also use the WoS subject categories in this paper.
We now turn to the three source normalization approaches that we study. In these approaches, a *source normalized citation score* (SNCS) is calculated for each publication. Since we have three source normalization approaches, we distinguish

between the $SNCS^{(1)}$, the $SNCS^{(2)}$, and the $SNCS^{(3)}$ value of a publication. The general idea of the three source normalization approaches is to weight each citation received by a publication based on the referencing behavior of the citing publication or the citing journal. The three source normalization approaches differ from each other in the exact way in which the weight of a citation is determined.

An important concept in the case of all three source normalization approaches is the notion of an active reference (Zitt & Small, 2008). In our analysis, an active reference is defined as a reference that falls within a certain reference window and that points to a publication in a WoS core journal. For instance, in the case of a four-year reference window, the number of active references in a publication from 2008 equals the number of references in this publication that point to publications in WoS core journals in the period 2005–2008. References to sources not covered by the WoS database or to WoS-indexed publications in non-core journals do not count as active references.

The $SNCS^{(1)}$ value of a publication is calculated as

$$SNCS^{(1)} = \sum_{i=1}^{c} \frac{1}{a_i} \tag{2}$$

where a_i denotes the average number of active references in all publications that appeared in the same journal and in the same year as the publication from which the ith citation originates. The length of the reference window within which active references are counted equals the length of the citation window of the publication for which the $SNCS^{(1)}$ value is calculated. The following example illustrates the definition of a_i. Suppose that we want to calculate the $SNCS^{(1)}$ value of a publication from 2008, and suppose that the ith citation received by this publication originates from a citing publication from 2010. Since the publication for which the $SNCS^{(1)}$ value is calculated has a four-year citation window (i.e., 2008–2011), a_i equals the average number of active references in all publications that appeared in the citing journal in 2010, where active references are counted within a four-year reference window (i.e., 2007–2010). The $SNCS^{(1)}$ approach is based on the idea of the audience factor of Zitt and Small (2008), although it applies this idea to an individual publication rather than an entire journal. Unlike the audience factor, the $SNCS^{(1)}$ approach uses multiple citing years.

The $SNCS^{(2)}$ approach is similar to the $SNCS^{(1)}$ approach, but instead of the average number of active references in a citing journal it looks at the number of active references in a citing publication. In mathematical terms,

$$SNCS^{(2)} = \sum_{i=1}^{c} \frac{1}{r_i} \tag{3}$$

where r_i denotes the number of active references in the publication from which the ith citation originates. Analogous to the $SNCS^{(1)}$ approach, the length of the

reference window within which active references are counted equals the length of the citation window of the publication for which the SNCS$^{(2)}$ value is calculated. The SNCS$^{(2)}$ approach is based on the idea of fractional citation counting of Leydesdorff and Opthof (2010; see also Leydesdorff & Bornmann, 2011; Leydesdorff et al., in press; Leydesdorff et al., 2013; Zhou & Leydesdorff, 2011). However, a difference with the fractional citation counting idea of Leydesdorff and Opthof is that instead of all references in a citing publication only active references are counted. This is a quite important difference. Counting all references rather than active references only disadvantages fields in which a relatively large share of the references point to older literature, to sources not covered by the WoS database, or to WoS-indexed publications in non-core journals.

The SNCS$^{(3)}$ approach, the third source normalization approach that we consider, combines ideas of the SNCS$^{(1)}$ and SNCS$^{(2)}$ approaches. The SNCS$^{(3)}$ value of a publication equals

$$\mathrm{SNCS}^{(3)} = \sum_{i=1}^{c} \frac{1}{p_i r_i} \tag{4}$$

where r_i is defined in the same way as in the SNCS$^{(2)}$ approach and where p_i denotes the proportion of publications with at least one active reference among all publications that appeared in the same journal and in the same year as the ith citing publication. Comparing (3) and (4), it can be seen that the SNCS$^{(3)}$ approach is identical to the SNCS$^{(2)}$ approach except that p_i has been added to the calculation. By including p_i, the SNCS$^{(3)}$ value of a publication depends not only on the referencing behavior of citing publications (like the SNCS$^{(2)}$ value) but also on the referencing behavior of citing journals (like the SNCS$^{(1)}$ value). The rationale for including p_i is that some fields have more publications without active references than others, which may distort the normalization implemented in the SNCS$^{(2)}$ approach. For a more extensive discussion of this issue, we refer to Waltman et al. (2013), who present a revised version of the SNIP indicator originally introduced by Moed (2010). The SNCS$^{(3)}$ approach is based on similar ideas as this revised SNIP indicator, although in the SNCS$^{(3)}$ approach these ideas are applied to individual publications while in the revised SNIP indicator they are applied to entire journals. Also, the SNCS$^{(3)}$ approach uses multiple citing years, while the revised SNIP indicator uses a single citing year.

Results

We split the discussion of the results of our analysis in two subsections. In the first subsection, we present results that were obtained by using the WoS journal subject categories to evaluate the normalization approaches introduced in the previous section. We then argue that this way of evaluating the different normalization approaches is likely to produce biased results. In the second

subsection, we use our algorithmically constructed classification systems A, B, and C instead of the WoS subject categories. We argue that this yields a fairer comparison of the different normalization approaches.

Results based on the Web of Science journal subject categories

Before presenting our results, we need to discuss how publications belonging to multiple WoS subject categories were handled. In the approach that we have taken, each publication is fully assigned to each of the subject categories to which it belongs. This means that some publications occur multiple times in the analysis, once for each of the subject categories to which they belong. Because of this, the total number of publications in the analysis is 6.47 million.

Table 2 reports for each year in the period 2007–2010 the average normalized citation score of all publications from that year, where normalized citation scores have been calculated using each of the four normalization approaches introduced in the previous section. The average citation score (CS) without normalization is reported as well. As expected, unnormalized citation scores display a decreasing trend over time. This can be explained by the lack of a correction for the age of publications. Table 2 also lists the number of publications per year. Notice that each year the number of publications is 3% to 5% larger than the year before.

Table 2. Average normalized citation score per year calculated using four different normalization approaches and the unnormalized CS approach. The citation scores are based on the 6.47 million publications included in the WoS journal subject categories classification system.

	2007	2008	2009	2010
No. of publications	1.51M	1.59M	1.66M	1.71M
CS	10.78	8.16	5.50	2.70
NCS	1.01	1.01	1.02	1.02
$SNCS^{(1)}$	1.10	1.07	1.07	1.05
$SNCS^{(2)}$	1.03	0.97	0.89	0.68
$SNCS^{(3)}$	1.10	1.07	1.07	1.05

Based on Table 2, we make the following observations:

- Each year, the average NCS value is slightly above one. This is a consequence of the fact that publications belonging to multiple subject categories are counted multiple times. Average NCS values of exactly one would have been obtained if there had been no publications that belong to more than one subject category.
- The average $SNCS^{(2)}$ value decreases considerably over time. The value in 2010 is more than 30% lower than the value in 2007. This shows that the $SNCS^{(2)}$ approach fails to properly correct for the age of a publication. Recent publications have a significant disadvantage compared with older ones. This is caused by the fact that in the $SNCS^{(2)}$ approach publications without active references give no 'credits' to earlier publications (see also

Waltman & Van Eck, in press; Waltman et al., 2013). In this way, the balance between publications that provide credits and publications that receive credits is distorted. This problem is most serious for recent publications. In the case of recent publications, the citation and reference windows used in the calculation of $SNCS^{(2)}$ values are relatively short, and the shorter the length of the reference window within which active references are counted, the larger the number of publications without active references.

- The $SNCS^{(1)}$ and $SNCS^{(3)}$ approaches yield the same average values per year. These values are between 5% and 10% above one (see also Waltman & Van Eck, in press), with a small decreasing trend over time. Average $SNCS^{(1)}$ and $SNCS^{(3)}$ values very close to one would have been obtained if there had been no increase in the yearly number of publications (for more details, see Waltman & Van Eck, 2010; Waltman et al., 2013). The sensitivity of source normalization approaches to the growth rate of the scientific literature was already pointed out by Zitt and Small (2008).

Table 2 provides some insight into the degree to which the different normalization approaches succeed in correcting for the age of publications. However, the table does not show to what extent each of the normalization approaches manages to correct for differences in citation practices between scientific fields. This raises the question when exactly we can say that differences in citation practices between fields have been corrected for. With respect to this question, we follow a number of recent papers (Crespo, Herranz, et al., 2012; Crespo, Li, et al., 2012; Radicchi & Castellano, 2012a, 2012c; Radicchi et al., 2008). In line with these papers, we say that the degree to which differences in citation practices between fields have been corrected for is indicated by the degree to which the normalized citation distributions of different fields coincide with each other. Differences in citation practices between fields have been perfectly corrected for if, after normalization, each field is characterized by exactly the same citation distribution. Notice that correcting for the age of publications can be defined in an analogous way. We therefore say that publication age has been corrected for if different publication years are characterized by the same normalized citation distribution.

The next question is how the similarity of citation distributions can best be assessed. To address this question, we follow an approach that was recently introduced by Crespo, Herranz, et al. (2012) and Crespo, Li, et al. (2012). For each of the four normalization approaches that we study, we take the following steps:

1. Calculate each publication's normalized citation score.
2. For each combination of a publication year and a subject category, assign publications to quantile intervals based on their normalized citation score. We work with 100 quantile (or percentile) intervals. Publications are sorted in ascending order of their normalized citation score, and the first 1% of the publications are assigned to the first quantile interval, the next

1% of the publications are assigned to the second quantile interval, and so on.

3. For each combination of a publication year, a subject category, and a quantile interval, calculate the number of publications and the average normalized citation score per publication. We use $n(q, i, j)$ and $\mu(q, i, j)$ to denote, respectively, the number of publications and the average normalized citation score for publication year i, subject category j, and quantile interval q.
4. For each quantile interval, determine the degree to which publication age and differences in citation practices between fields have been corrected for. To do so, we calculate for each quantile interval q the inequality index $I(q)$ defined by

$$I(q) = \frac{1}{n(q)} \sum_{i=2007}^{2010} \sum_{j=1}^{m} n(q,i,j) \frac{\mu(q,i,j)}{\mu(q)} \log\left(\frac{\mu(q,i,j)}{\mu(q)} \right) \quad (5)$$

where m denotes the number of subject categories, $n(q)$ denotes the number of publications in quantile interval q aggregated over all publication years and subject categories, and $\mu(q)$ denotes the average normalized citation score of these publications. The inequality index $I(q)$ in (5) is known as the Theil index. We refer to Crespo, Li, et al. (2012) for a justification for the use of this index. The lower the value of the index, the better the correction for publication age and field differences. A perfect normalization approach would result in $I(q) = 0$ for each quantile interval q. In the calculation of $I(q)$ in (5), we use natural logarithms and we define $0 \log(0) = 0$. Notice that $I(q)$ is not defined if $\mu(q) = 0$.

We perform the above steps for each of our four normalization approaches. Moreover, for the purpose of comparison, we perform the same steps also for citation scores without normalization.

The results of the above calculations are presented in Figure 1. For each of our four normalization approaches, the figure shows the value of $I(q)$ for each of the 100 quantile intervals. For comparison, $I(q)$ values calculated based on unnormalized citation scores are displayed as well. Notice that the vertical axis in Figure 1 has a logarithmic scale.

As expected, Figure 1 shows that all four normalization approaches yield better results than the approach based on unnormalized citation scores. For all or almost all quantile intervals, the latter approach, referred to as the CS approach in Figure 1, yields the highest $I(q)$ values. It can further be seen that the NCS approach significantly outperforms all three SNCS approaches. Hence, in line with recent studies by Leydesdorff et al. (in press) and Radicchi and Castellano (2012a), Figure 1 suggests that classification-system-based normalization is more accurate than source normalization. Comparing the different SNCS approaches, we see that

the SNCS$^{(2)}$ approach is outperformed by the SNCS$^{(1)}$ and SNCS$^{(3)}$ approaches. Notice further that for all normalization approaches $I(q)$ values are highest for the lowest quantile intervals. These quantile intervals include many uncited and very lowly cited publications. From the point of view of the normalization of citation scores, these quantile intervals may be considered of less interest, and it may be best to focus mainly on the higher quantile intervals.

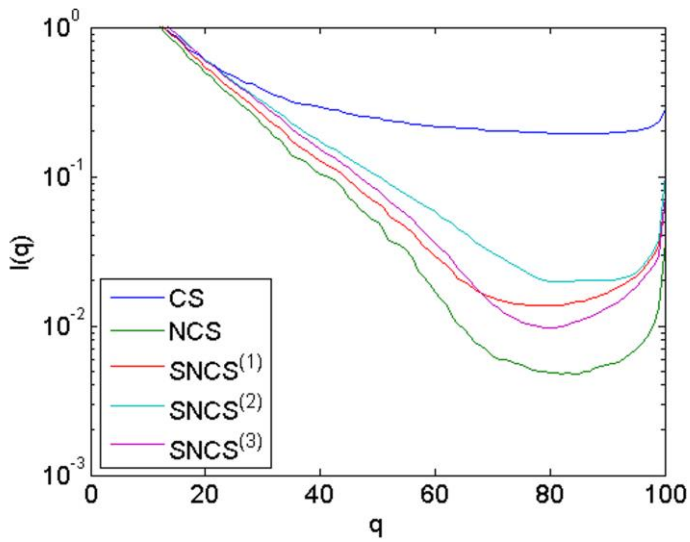


Figure 1. Inequality index $I(q)$ calculated for 100 quantile intervals q and for four different normalization approaches. Results calculated for the unnormalized CS approach are displayed as well. All results are based on the WoS journal subject categories classification system.

The above results may seem to provide clear evidence for preferring classification-system-based normalization over source normalization. However, there may be a bias in the results that causes the NCS approach to have an unfair advantage over the three SNCS approaches. The problem is that the WoS subject categories are used not only in the evaluation of the different normalization approaches but also in the implementation of one of these approaches, namely the NCS approach. The standard used to evaluate the normalization approaches should be completely independent of the normalization approaches themselves, but for the NCS approach this is not the case. Because of this, the above results may be biased in favor of the NCS approach. In the next subsection, we therefore use our algorithmically constructed classification systems A, B, and C to evaluate the different normalization approaches in a fairer way.

Before proceeding to the next subsection, we note that the above-mentioned studies by Leydesdorff et al. (in press) and Radicchi and Castellano (2012a) suffer from the same problem as our above results. In these studies, the same

classification system is used both in the implementation and in the evaluation of a classification-system-based normalization approach. This is likely to introduce a bias in favor of this normalization approach. This problem was first pointed out by Sirtes (2012) in a comment on Radicchi and Castellano's (2012a) study (for the rejoinder, see Radicchi & Castellano, 2012b).

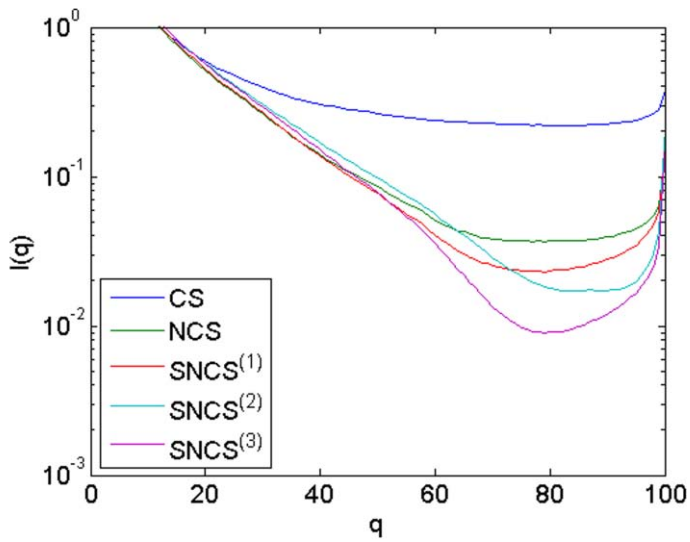


Figure 2. Inequality index *I*(*q*) calculated for 100 quantile intervals *q* and for four different normalization approaches. Results calculated for the unnormalized CS approach are displayed as well. All results are based on classification system C.

Results based on classification systems A, B, and C

We now present the results obtained by using the algorithmically constructed classification systems A, B, and C to evaluate the four normalization approaches that we study. As we have argued above, this yields a fairer comparison of the different normalization approaches than an evaluation using the WoS subject categories. In classification systems A, B, and C, each publication belongs to only one research area.

We examine the degree to which, after applying one of our four normalization approaches, different fields and different publication years are characterized by the same citation distribution. To assess the similarity of citation distributions, we take the same steps as described in the previous subsection, but with fields defined by research areas in our classification systems A, B, and C rather than by WoS subject categories. The results obtained for classification system C are shown in Figure 2. Due to space limitations, the results obtained for classification systems A and B are not shown. However, these results can be found in Figures 2 and 3 in the extended version of this paper (Waltman & Van Eck, 2013).

The following observations can be made based on Figure 2 combined with Figures 2 and 3 in the extended version of this paper:

- Like in Figure 1, the CS approach, which does not involve any normalization, is outperformed by all four normalization approaches.
- The results presented in Figure 1 are indeed biased in favor of the NCS approach. Compared with Figure 1, the performance of the NCS approach in Figure 2 is disappointing. In the case of classification systems B and C, the NCS approach is significantly outperformed by both the $SNCS^{(1)}$ and the $SNCS^{(3)}$ approach. In the case of classification system A, the NCS approach performs better, although it is still outperformed by the $SNCS^{(1)}$ approach.
- Like in Figure 1, the $SNCS^{(2)}$ approach is consistently outperformed by the $SNCS^{(3)}$ approach. In the case of classification systems A and B, the $SNCS^{(2)}$ approach is also outperformed by the $SNCS^{(1)}$ approach. It is clear that the disappointing performance of the $SNCS^{(2)}$ approach must at least partly be due to the failure of this approach to properly correct for publication age, as we have already seen in Table 2.
- The $SNCS^{(1)}$ approach has a mixed performance. It performs very well in the case of classification system A, but not so well in the case of classification system C. The $SNCS^{(3)}$ approach, on the other hand, has a very good performance in the case of classification systems B and C, but this approach is outperformed by the $SNCS^{(1)}$ approach in the case of classification system A.

The overall conclusion is that in order to obtain the most accurate normalized citation scores one should generally use a source normalization approach rather than a normalization approach based on the WoS subject categories classification system. However, consistent with our earlier work (Waltman & Van Eck, in press), it can be concluded that the $SNCS^{(2)}$ approach should not be used. Furthermore, the $SNCS^{(3)}$ approach appears to be preferable over the $SNCS^{(1)}$ approach. The excellent performance of the $SNCS^{(3)}$ approach in the case of classification system C (see Figure 2) suggests that this approach is especially well suited for fine-grained analyses aimed for instance at comparing researchers or research groups active in different subfields within the same field.

Some more detailed results are presented in Appendix C in the more extensive version of this paper (Waltman & Van Eck, 2013). In this appendix, we use a decomposition of citation inequality proposed by Crespo, Herranz, et al. (2012) and Crespo, Li, et al. (2012) to summarize in a single number the degree to which each of our normalization approaches has managed to correct for differences in citation practices between fields and differences in the age of publications.

Conclusions

In this paper, we have addressed the question how citation-based bibliometric indicators can best be normalized to ensure fair comparisons between publications from different scientific fields and different years. In a systematic large-scale

empirical analysis, we have compared a normalization approach based on a field classification system with three source normalization approaches. In the classification-system-based normalization approach, we have used the WoS journal subject categories to classify publications into fields. The three source normalization approaches are inspired by the audience factor of Zitt and Small (2008), the idea of fractional citation counting of Leydesdorff and Opthof (2010), and our own revised SNIP indicator (Waltman et al., 2013).

Compared with earlier studies, our analysis offers three methodological innovations. Most importantly, we have distinguished between the use of a field classification system in the implementation and in the evaluation of a normalization approach. Following Sirtes (2012), we have argued that the classification system used in the evaluation of a normalization approach should be different from the one used in the implementation of the normalization approach. We have demonstrated empirically that the use of the same classification system in both the implementation and the evaluation of a normalization approach leads to significantly biased results. Building on our earlier work (Waltman & Van Eck, in press), another methodological innovation is the exclusion of special types of publications, for instance publications in national scientific journals, popular scientific magazines, and trade magazines. A third methodological innovation is the evaluation of normalization approaches at different levels of granularity. As we have shown, some normalization approaches perform better at one level than at another.

Based on our empirical results and in line with our earlier work (Waltman & Van Eck, in press), we advise against using source normalization approaches that follow the fractional citation counting idea of Leydesdorff and Opthof (2010). The fractional citation counting idea does not offer a completely satisfactory normalization (see also Waltman et al., 2013). In particular, we have shown that it fails to properly correct for the age of a publication. The other two source normalization approaches that we have studied generally perform better than the classification-system-based normalization approach based on the WoS subject categories, especially at higher levels of granularity. It may be that other classification-system-based normalization approaches, for instance based on algorithmically constructed classification systems, have a better performance than subject-category-based normalization. However, any classification system can be expected to introduce certain biases in a normalization, simply because any organization of the scientific literature into a number of perfectly separated fields of science is artificial. So consistent with our previous study (Waltman & Van Eck, in press), we recommend the use of a source normalization approach. Except at very low levels of granularity (e.g., comparisons between broad disciplines), the approach based on our revised SNIP indicator (Waltman et al., 2013) turns out to be more accurate than the approach based on the audience factor of Zitt and Small (2008). Of course, when using a source normalization approach, it should always be kept in mind that there are certain factors, such as the growth rate of the scientific literature, for which no correction is made.

References

Crespo, J.A., Herranz, N., Li, Y., & Ruiz-Castillo, J. (2012). *Field normalization at different aggregation levels* (Working Paper Economic Series 12-22). Departamento de Economía, Universidad Carlos III of Madrid.

Crespo, J.A., Li, Y., & Ruiz-Castillo, J. (2012). *Differences in citation impact across scientific fields* (Working Paper Economic Series 12-06). Departamento de Economía, Universidad Carlos III of Madrid.

Glänzel, W., Schubert, A., Thijs, B., & Debackere, K. (2011). A priori vs. a posteriori normalisation of citation indicators. The case of journal ranking. *Scientometrics*, *87*(2), 415–424.

Leydesdorff, L., & Bornmann, L. (2011). How fractional counting of citations affects the impact factor: Normalization in terms of differences in citation potentials among fields of science. *Journal of the American Society for Information Science and Technology*, *62*(2), 217–229.

Leydesdorff, L., & Opthof, T. (2010). Scopus's source normalized impact per paper (SNIP) versus a journal impact factor based on fractional counting of citations. *Journal of the American Society for Information Science and Technology*, *61*(11), 2365–2369.

Leydesdorff, L., Radicchi, F., Bornmann, L., Castellano, C., & De Nooy, W. (in press). Field-normalized impact factors: A comparison of rescaling versus fractionally counted IFs. *Journal of the American Society for Information Science and Technology*.

Leydesdorff, L., Zhou, P., & Bornmann, L. (2013). How can journal impact factors be normalized across fields of science? An assessment in terms of percentile ranks and fractional counts. *Journal of the American Society for Information Science and Technology*, *64*(1), 96–107.

Moed, H.F. (2010). Measuring contextual citation impact of scientific journals. *Journal of Informetrics*, *4*(3), 265–277.

Radicchi, F., & Castellano, C. (2012a). Testing the fairness of citation indicators for comparison across scientific domains: The case of fractional citation counts. *Journal of Informetrics*, *6*(1), 121–130.

Radicchi, F., & Castellano, C. (2012b). Why Sirtes's claims (Sirtes, 2012) do not square with reality. *Journal of Informetrics*, *6*(4), 615–618.

Radicchi, F., & Castellano, C. (2012c). A reverse engineering approach to the suppression of citation biases reveals universal properties of citation distributions. *PLoS ONE*, *7*(3), e33833.

Radicchi, F., Fortunato, S., & Castellano, C. (2008). Universality of citation distributions: Toward an objective measure of scientific impact. *Proceedings of the National Academy of Sciences*, *105*(45), 17268–17272.

Sirtes, D. (2012). Finding the Easter eggs hidden by oneself: Why Radicchi and Castellano's (2012) fairness test for citation indicators is not fair. *Journal of Informetrics*, *6*(3), 448–450.

Van Eck, N.J., Waltman, L., Van Raan, A.F.J., Klautz, R.J.M., & Peul, W.C. (2012). *Citation analysis may severely underestimate the impact of clinical research as compared to basic research*. arXiv:1210.0442.

Waltman, L., & Van Eck, N.J. (2010). The relation between Eigenfactor, audience factor, and influence weight. *Journal of the American Society for Information Science and Technology*, *61*(7), 1476–1486.

Waltman, L., & Van Eck, N.J. (2012). A new methodology for constructing a publication-level classification system of science. *Journal of the American Society for Information Science and Technology*, *63*(12), 2378–2392.

Waltman, L., & Van Eck, N.J. (2013). *A systematic empirical comparison of different approaches for normalizing citation impact indicators*. arXiv:1301.4941.

Waltman, L., & Van Eck, N.J. (in press). Source normalized indicators of citation impact: An overview of different approaches and an empirical comparison. *Scientometrics*.

Waltman, L., Van Eck, N.J., Van Leeuwen, T.N., & Visser, M.S. (2013). Some modifications to the SNIP journal impact indicator. *Journal of Informetrics*, *7*(2), 272–285.

Zhou, P., & Leydesdorff, L. (2011). Fractional counting of citations in research evaluation: A cross- and interdisciplinary assessment of the Tsinghua University in Beijing. *Journal of Informetrics*, *5*(3), 360–368.

Zitt, M. (2010). Citing-side normalization of journal impact: A robust variant of the audience factor. *Journal of Informetrics*, *4*(3), 392–406.

Zitt, M. (2011). Behind citing-side normalization of citations: Some properties of the journal impact factor. *Scientometrics*, *89*(1), 329–344.

Zitt, M., & Small, H. (2008). Modifying the journal impact factor by fractional citation weighting: The audience factor. *Journal of the American Society for Information Science and Technology*, *59*(11), 1856–1860.

THE TIPPING POINT – OPEN ACCESS COMES OF AGE

Éric Archambault

Eric.archambault@science-metrix.com
Science-Metrix Inc., Observatoire des Sciences et des Technologies (OST),
Centre Interuniversitaire de Recherche sur la Science et la Technologie (CIRST),
Université du Québec à Montréal, Montréal, Québec, Canada

Abstract
The Open Access (OA) model for scientific publications has been examined for years by academics who have argued that it presents advantages in increasing accessibility and, consequently, in increasing the impact of papers. It has been noted that OA availability has increased steadily over the years. However, current measurement has seriously underestimated the proportion of OA peer-reviewed articles. This paper presents the results of a pilot study that shows evidence that the proportion of measured OA is so close to 50% that we have most likely passed the tipping point, that is, the stage where the majority of articles become available for free.

Conference Topic
Topic 10**:** Open Access and Scientometrics

Introduction
Interest in the academic community for Open Access (OA) publications has been increasing. The initial interest in the use of bibliometric methods focused on accessing the so-called citation advantage of OA as opposed to subscription-based journals (Antelman, 2004; Harnad & Brody, 2004; Craig, 2007). The literature of the time recognised a clear citation advantage to papers available in OA as opposed to papers diffused solely through subscription-based journals. Strong advocacy by authors such as Harnad (2003, 2008, 2012) suggested that benefits would ensue from so-called green OA, that is, research papers self-archived by their authors in various types of repositories. Unsurprisingly, in this context, librarians and information scientists noted that they had a new mission, which meant setting up and curating OA repositories (Proser, 2003; Bailey, 2005; Chan, Kwok, & Yip, 2005; Chan, Devakos & Mircea, 2005 Repanovici, 2012).
A part of the OA literature has discussed how authors and researchers (Pelizzari, 2004; Swan & Brown, 2004; Dubini, Galimberti & Micheli, 2010) and publishers (Morris, 2003; Regazzi, 2004) would react to this new paradigm. Evidently, business and economic models were discussed (Bilder, 2003; Kurek, Geurts & Roosendaal, 2006; Houghton, 2010; Lakshmi Poorna, Mymoon & Hariharan, 2012), but there was also interest in what models academia and libraries would follow (Rowland et al., 2004; Swan et al., 2005; Hu, Zhang & Chen, 2010).

As OA continued to make inroads, a growing number of papers examined the state of development of OA in specific countries (Nyambi & Maynard, 2012; Sawant, 2012; Woutersen-Windhouwer, 2012; Miguel et al., 2013) and in specific fields of research (Abad-Garcí et al., 2010; Gentil-Beccot, Mele, & Brook, 2010; Charles, & Booth, 2011; Henderson, 2013). In this context, it was not surprising to find papers that addressed the general question of OA availability as a proportion of the scientific literature, and the proportion of OA papers available in different fields of science (Björk et al. 2010; Gargouri et al., 2012).
This paper re-assesses OA availability in 2008 through a careful examination of recall, which leads to a doubling of the proportion of OA estimated by Björk et al. and by Gargouri et al. The paper argues that the tipping point for OA has been reached and that one can expect that, from the late 2000s onwards, the majority of published academic peer-reviewed journal articles were available for free to end-users. The paper presents data for 22 fields of science as well as for the European Research Area countries, Brazil, Canada, Japan, and the US.

Methods

Accuracy and Precision: The paper presents the results for the pilot phase of a study that aims to estimate the *proportion of peer-reviewed journal articles which are freely available, that is, OA* for the last ten years (the pilot study is on OA availability in 2008). It builds on two important concepts: (1) *accuracy*, reflected in the quality of the instruments used and the care taken in making measurements; (2) *precision*, which involves repeated measures, sampling and statistical analysis (see figure 1)—the later concept will be called *statistical precision* for reasons that will become obvious.

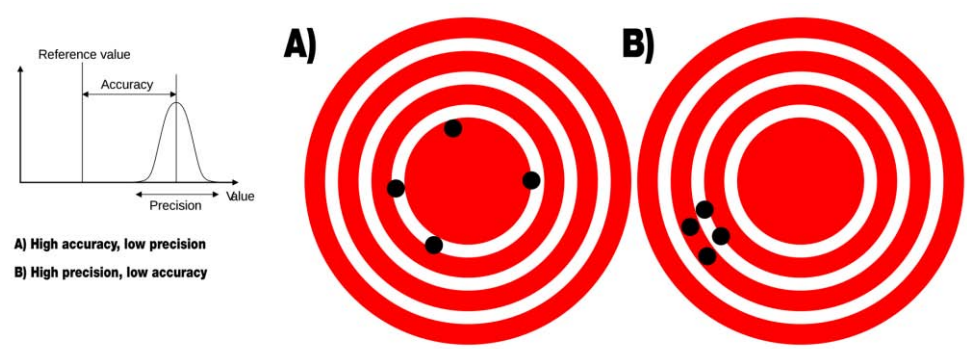


Figure 1. Accuracy and statistical precision (Adapted from http://en.wikipedia.org/wiki/Accuracy_and_precision)

Statistical precision can be approximated with the margin of error (ME). For a proportion (p) where the population is finite and known (N), is not systematically much larger than the sample size (n), and in which the values are discrete (for

example, papers), given a critical score Z (which will be set at 0.95 in the study), ME is calculated as follows:

$$ME = Z\sqrt{\frac{p(1-p)(N-n)}{n(N-1)}} + \frac{0.5}{n}$$

What complicates the use of these definitions is the need to examine accuracy with two more concepts used in information retrieval: recall and precision (hence the need to call the previous concept 'statistical precision'; the second precision-related concept will be knonw as 'retrieval precision'). Recall is the proportion of relevant records that are retrieved, while retrieval precision is the proportion of retrieved records that are relevant. If an instrument retrieves 25 records of which only 20 are relevant, and fails to retrieve 30 additional relevant records, its retrieval precision is 20/25 = 80% while its recall is 25/50 = 50%. Precision is synonymous with Type I errors (false positives), and recall with type II errors (false negatives). Thus, a high recall means that an instrument returned most of the relevant results, while high retrieval precision means that it retrieved more relevant results than irrelevant ones. Note that assessing the real positives accurately is frequently a distinct problem, as is the case in the present study.

Let us call π the proportion of the whole population of peer-reviewed papers that are OA. One cannot easily measure π directly because the population of scientific papers is relatively large, and there is currently no satisfactory complete repertory of that population. Hence, it is unlikely in the short term that someone will find another way than sampling to calculate p, an approximation of π. Though it is nearly impossible for p to equal π, it is the aim of this study to offer a robust design that will ensure that p is reasonably close to π. In the present study, two principal proportions will be calculated: (1) the overall proportion of OA literature; and (2) the proportion of the scientific literature published in gold journals. Before entering into the methodological details associated with the measurement as such, it is important to produce operational definitions of OA, green OA, gold OA and hybrid OA.

Types of OA scientific literature: Peter Suber suggests that '[o]pen-access (OA) literature is digital, online, free of charge, and free of most copyright and licensing restrictions.'[164] An effective definition of OA for this study is the following: 'OA, whether Green or Gold, is about giving people free access to peer-reviewed research journal articles.'[165] The following operational definitions of gold and green OA will be used in the present study.

- Gold OA refers to papers published in journals that provide free access to [peer-reviewed scholarly] papers. Authors sometimes, but not always, pay a fee for these publications. In the present study, Gold journals are those that provide cover-to-cover, instant access to articles.

[164] http://www.earlham.edu/~peters/fos/overview.htm.

[165] http://scholarlykitchen.sspnet.org/2011/09/07/oa-rhetoric-economics-and-the-definition-of-research/.

- Green OA generally refers to authors' self-archiving [of papers accepted in academic journals following a successful peer-review process].
- Hybrid OA is an increasingly important trend in scientific publishing by which authors pay for their papers to be available in OA in an otherwise not OA journal—'[h]ybrid open access journals provide Gold OA only for those individual articles for which their authors (or their author's institution or funder) pay an OA publishing fee.'[166]

A note on the concept of open-access versus toll-access literature is in order here. OA is rarely free, and can generally be seen as moving the toll plaza before the publication process as opposed to placing it after it. Open access will rarely entirely miss exacting a toll somewhere, be it on taxpayers' or on philanthropists' funds, or on the time of volunteers. Thus, the term toll-access, to distinguish the non-OA literature, is avoided here.

Peer-reviewed journal articles and original contributions to knowledge: A central part of the scientific literature is comprised of papers published in peer-reviewed journals (Larivière et al., 2006). This study concentrates on peer-reviewed, scholarly articles and omits the many other types of vehicles that are used for the *written* diffusion of scientific knowledge, namely books and conference proceedings, as well as research reports, mimeos and other heterogeneous forms, collectively called grey literature. A best practice in bibliometrics is to use only articles that can be considered original contributions to knowledge. The tradition in the Web of Knowledge (and its predecessor, the Science Citation Index) was to restrict the selection of document types to articles, notes and reviews (Carpenter & Narin, 1980). In Scopus, the tagging of articles is substantially more complex, and a combination of Source Type and Document Type is required to keep only what can be considered original contributions to knowledge. The present study uses the following operational definition: *articles that use references and are cited*. This definition, and empirically obtained thresholds, can be used to prune the Scopus production database of trade journals and non-original contributions to knowledge (at the macro level rather than at the article level to prevent the exclusion of papers that have not yet been cited). The resulting types of documents used are presented in the accompanying side box.

Source Type	Document Type
Book Series	Article
	Conference Paper
	Review
	Short Survey
Conference Proceeding	Article
	Review
Journal	Article
	Conference Paper
	Review
	Short Survey

Calculating the denominator: An important aspect of the project involves determining the proportion of OA papers by precisely estimating the number of

[166] http://en.wikipedia.org/wiki/Open_access.

OA peer-reviewed papers (the numerator) and dividing this by a carefully designed estimate of the number of peer-reviewed articles (the denominator) for each of the selected 22 disciplines and for the total literature. A decision was made to use the Ulrich periodical database to provide an estimate of the denominator, and these data and rights to use them were acquired for this study. The strengths and weaknesses of Ulrich data are well known: for example, some journals that should be classified as peer-reviewed are not (and the reverse is also true). A good example of this is the OA journal *Activités*, which mentions that 'Texts that have been submitted to *Activités* (www.activites.org/) will be assessed by two referees (called upon in view of the article). Each will give his or her opinion on the text.'[167] Despite this, and a description that clearly suggests scholarly content and the presence in papers of references to scholarly work, Ulrich has not classified this journal as refereed. Although several journals are likely to be classified 'Academic/Scholarly' in Ulrich and might be considered as contributing to science, this category cannot be included *en masse* as it comprises a substantial amount of material published in universities that has little scientific content. This is the case, for example, with the 'The Hilltop', classified by Ulrich as Academic/Scholarly, and claiming to be the 'The Student Voice of Howard University' (see http://www.thehilltoponline.com/). Consequently, the selection was restricted to Ulrich listed *journals* considered *refereed/peer-reviewed* AND Academic/Scholarly. Although imperfect, Ulrich remains the most extensive and authoritative and probably the least biased source of data on academic peer-reviewed journals and is therefore a solid calibration instrument for a systematic investigation of the peer-reviewed literature.

The core use of Ulrich in this project was to calibrate the proportion of papers from each of 22 disciplines used to present disaggregated statistics. The reason Ulrich is preferred is because *article-level* database publishers such as Elsevier (publisher of Scopus) and Thomson (Web of Science) are faced with choices having important commercial and profitability impacts. When selecting journals to be included for an article-level database such as Scopus, deciding whether to include a journal has a direct impact on production costs and partly because of this, database publishers tend to have a bias towards larger journals (economies of scale) and larger publishers (lowest transaction costs and economies of scale). However, whether a journal is small or large in terms of number of articles has substantially fewer consequences when it is included in a journal title database, where journal size can be expected to little impact on cost (some differences remain as it is likely easier to find information about the larger journals).

Ulrich cannot be used alone as it does not contain article-level information. The core work of the present project involved using a fully-licensed version of Elsevier's Scopus database hosted in house and conditioned over several years to produce bibliometric statistics. This meant that it was possible to randomly select papers among the millions of papers indexed. Ulrich was used to 'calibrate' the

[167] http://www.activites.org/resources/activites.eng.book.pdf.

proportion of peer-reviewed journal articles for each of the 22 fields used in this paper to present detailed statistics. The technique used to determine this proportion involved the following steps: 1) journals in Ulrich were matched to those contained in Scopus; 2) journals that intersected were given the discipline that was already contained in our classification of Scopus journals (for those that did not intersect, the Ulrich classification was compared with that used in our classification, and a matching table was used to attribute one of 22 disciplines to each of the journals); and 3) the number of articles per discipline was counted in the intersecting set, while the number of articles in the Ulrich set with no Scopus counterparts was determined by projecting the average number of articles for the 50% journals in Scopus with the fewest articles per journal. The reason for using the average number of articles for the 50% smaller journals is that experience has revealed that databases such as Scopus and the Web of Science index the largest journals first. For instance, the Web of Science covers about 12,000 journals, and Scopus about 18,000. Despite a 50% increase in journal coverage, Scopus only has about 20% more articles. A sensitivity analysis was performed to see the effect of calculating the average for the 75%, 50%, and 25% smallest journals (ranked by decreasing number of articles), and the results were broadly similar.

Strategy to measure the proportion of gold OA: Somewhat distinct strategies were used to calculate the occurrence of gold OA and total OA. For gold articles, an estimate of the proportion of papers was made from the random sample by matching the journals that were known to be gold in 2008. These journals were obtained from the Directory of Open Access Journals (DOAJ) and the list of OA journals in PubMed Central. This was done by matching journals' ISSN, E-ISSN and names from Scopus to the relevant records in the sample (the matching had about 100% precision, but recall may have been imperfect, hence the figures presented here can be considered a floor, rather than a ceiling).

Strategy to measure the OA proportion of scientific articles: Two samples and a sub-sample were produced to undertake a pilot study to measure OA availability in 2008 given the definition and assumptions presented above. A first sample of 20,000 was produced for early testing, and a sub-sample of 500 records was drawn from this sample to determine the availability of papers in OA using various search engines; a 'ground truth' was established by combining the validated results of these tests.
A second random sample of 20,000 records was drawn from Scopus and used to perform the measuring stage of the pilot study. This sample was restricted to papers published in 2008, and the results were restricted to original contributions to knowledge; records where the journal name or the record type contained a conference were excluded. Records for which the discipline was unknown were also set aside. The eligible record set from 2008, comprising somewhat more than 1.36 million records in Scopus, was 'tossed' five times using a pseudo-random method (using the newid() command in SQL Server), a subset of 100,000 records

was selected, placed in a subset, and tossed again. These 100,000 records were then imported into Excel, where a straightforward analysis of the distribution of the records by discipline was performed. This analysis showed that a subsample of 20,000 records would keep few records in three of the smaller disciplines (Philosophy & Theology, Visual & Performing Arts, and General Arts, Humanities & Social Sciences). For these disciplines, a random sample of 100 records was selected, and for the Built Environment & Design discipline, the 101 records that were part of the 100,000 records were all selected. As the objective was to produce a record set of 20,000, a subsequent selection was done for 19,599 records. These were selected by tossing the 100,000 a few more times using the rand() command in Excel, then proceeding to the selection of the required number of records.

Technique used to harvest OA articles: Although the pilot study was meant to build on the method pioneered by Björk *et al*. and human judgment was to be used in searching for and categorising the presence of OA, the pilot study has led to a gradual, but fundamental, modification of the original approach. After nearly two months of work, it became apparent that using professionals would be cost-prohibitive and too slow for a large scale study over several years.

A test was then conducted with 20,000 records being provided to the Steven Harnad team in Montreal. This relatively blind test produced recall that was good; the scores computed were much higher than those presented in previous papers, including results by Harnad's team. This was due to the use of Scopus, as opposed to the Web of Science as Harnad's team had done before. Some 500 records of this set were then extracted randomly and extensive testing was performed. The records were all searched manually in Google Scholar, Google, and Microsoft Academics. Records that could be downloaded for free and that came from any of these sources were considered OA, and the carefully verified sample a 'ground truth.'

These tests led to the following observations: Google Scholar and Google have substantial overlap, but each search engine has a somewhat distinct set of positive results. Microsoft Academics does not add much to the combined results of Google and Google Scholar. Importantly also, the results obtained suggest that the accuracy of the harvesting instrument, and the coverage of the database, are more important than a large sample size (statistical precision). For instance, the team led by Harnad measured only 22% of OA in 2008 overall 'out of the 12,500 journals indexed by Thomson Reuters using a robot that trawled the Web for OA full-texts' (Gargouri et al., 2012). Likewise, Bj rk *et al*. found a score of 20% using Scopus and Google as a search engine. When the Harnad team ran their robot on our Scopus sample, the proportion of total OA jumped to close to 32%, compared with the 22% they obtained in WoS as mentioned in their paper (this original sample was prepared rapidly for testing and might not have been perfectly random, so these results should be seen as tentative). This shows that a

technique to measure the proportion of OA literature based on the Web of Science produces fairly low recall and seriously underestimates OA availability.
Extensive testing was done with the subsample of 500 records. Because the original was not necessarily 100% random, this subsample cannot necessarily be considered as totally representative, but the results are nonetheless instructive. The results for the Harnad robot are as is and contain a few false positives, so the real positive score is actually lower. The Scholar, Google and Ground Truth results were manually validated and the documents downloaded, and as such, they can be considered accurate. The Ground Truth comprises the combined validated results from Google and Google Scholar in addition to one result from Microsoft Academics. Results from Microsoft Academics are not shown, as only the negative results from Scholar and Google were tested to examine whether this added any substantial results to the previous ones.

Table 1. Availability of OA in a sample of 500 Scopus records, 2008

Result	UQAM (Harnad)	Scholar	Google	Ground Truth
FALSE	350	293	290	262
TRUE	150	207	210	238
Total	**500**	**500**	**500**	**500**
% OA	30%	41%	42%	48%

Source: Computed by Science-Metrix

This extensive analysis therefore suggests that 48% of the literature published in 2008 may be available for free. Despite their high level of performance, neither Google nor Google Scholar can be expected to crawl the Web perfectly or to have a search engine so robust that it systematically presents all the relevant records in the first page of results (which we limited our analysis to), and hence cannot be expected to have a 100% recall, especially for academic articles (Arlitsch & O'Brien, 2012). Consequently, one can infer that OA availability very likely passed the tipping point in 2008 (or earlier) and that the majority of peer-reviewed/scholarly papers published in journals in that year are now available for free in one form or another to end-users.
These results suggest that using Scopus and an improved harvester 'to trawl the Web for OA full-texts' could yield substantially more accurate results than the methods used by Björk *et al*. and Harnad *et al*.

Results

Table 1 presents data on OA availability overall and for Gold journals (pure Gold, in that it does not include journals with an embargo period or traditional-model journals offering pay-per-article OA). Pay-per-article OA, journals with embargo periods and journals allowing partial indexing following granting agencies' OA policies are considered hybrid, and these data are bundled here with green OA (self-archiving). Papers in each of the 22 fields have been recalibrated given the method presented before (calibration based on Ulrich).The overall rate calculated

with the current harvesting instrument is 42% (plus or minus three percentage points). Considering that the instrument used has imperfect recall and considering that OA Gold journals are likely to be under-represented in both Scopus and the Ulrich database, this can be considered a floor rather than an upper limit.
OA availability varies considerably among disciplines. It seems that the tipping point has been passed (OA availability over 50%) in Biology, Biomedical Research, Mathematics & Statistics, and General Science & Technology. According to these data, a third or less of the papers can be found in OA in Chemistry, Enabling & Strategic Technologies, Historical Studies, and Engineering, while less than one paper out of five can be accessed free in Communication & Textual Studies and in Visual & Performing Arts. However, one must be careful with these last two figures as the statistical error is of the same order of magnitude as the measured proportion.

Table 2.Proportion of OA per discipline, 2008

Field	Papers	Green & Hybrid		Gold		OA	
		Papers	%	Papers	%	Papers	%
Agriculture, Fisheries & Forestry	780	199	26 ± 6	125	16 ± 7	324	42 ± 4
Biology	1,031	477	46 ± 4	161	16 ± 6	638	62 ± 3
Biomedical Research	1,618	858	53 ± 2	141	9 ± 5	999	62 ± 2
Built Environment & Design	100	30	30 ± 15	7	7 ± 25	37	37 ± 14
Chemistry	1,621	379	23 ± 4	154	9 ± 5	532	33 ± 3
Clinical Medicine	5,157	1,609	31 ± 2	501	10 ± 2	2,110	41 ± 2
Communication & Textual Studies	249	33	13 ± 19	15	6 ± 24	48	19 ± 17
Earth & Environmental Sciences	599	228	38 ± 5	28	5 ± 9	256	43 ± 5
Economics & Business	627	246	39 ± 6	23	4 ± 11	269	43 ± 5
Enabling & Strategic Technologies	1,267	301	24 ± 4	75	6 ± 5	376	30 ± 4
Engineering	1,168	290	25 ± 4	17	1 ± 8	307	26 ± 4
General Arts, Humanities & Social Sciences	25	11	44 ± 12	0.2	1 ± 70	11	45 ± 12
General Science & Technology	165	52	32 ± 9	40	24 ± 10	92	56 ± 6
Historical Studies	232	48	21 ± 13	20	9 ± 17	68	29 ± 12
Information & Communication Technologies	590	220	37 ± 5	30	5 ± 9	250	42 ± 5
Mathematics & Statistics	625	333	53 ± 4	31	5 ± 10	364	58 ± 4
Philosophy & Theology	164	52	32 ± 15	10	6 ± 27	62	38 ± 14
Physics & Astronomy	1,872	747	40 ± 3	89	5 ± 5	836	45 ± 3
Psychology & Cognitive Sciences	436	193	44 ± 6	17	4 ± 13	210	48 ± 6
Public Health & Health Services	581	194	33 ± 6	70	12 ± 8	264	45 ± 5
Social Sciences	1,051	313	30 ± 6	96	9 ± 8	408	39 ± 5
Visual & Performing Arts	43	7	16 ± 20	0.9	2 ± 44	8	18 ± 19
All Publications	**20,000**	**6,818**	**34 ± 4**	**1,649**	**8 ± 6**	**8,467**	**42 ± 3**

It is more delicate to interpret the proportion of Gold OA because of the large statistical error (resulting from the small sample and low occurrence). The overall Gold OA availability measured here is 8%, and this is generally consistent with the literature. Note however that this report uses a strict definition of Gold OA, and that many previous studies might have included disembargoed papers and pay-per-article OA, which is not the case here. Gold OA is widespread in General Science & Technology, Agriculture, Fisheries & Forestry, Biology, Public Health & Health Services, and Clinical Medicine. Less than 2% of the papers are

available in Gold journals in Visual & Performing Arts, Engineering and General Arts, Humanities & Social Sciences.

The prevalence of papers in hybrid forms (non-Gold) is especially high in Mathematics & Statistics and Biomedical Research. Less than one paper out of four can be found in hybrid forms in Engineering, Enabling & Strategic Technologies, Chemistry, Historical Studies and the Visual & Performing Arts.

A question that has animated OA advocates has been the so-called citation advantage of OA. This question is examined briefly in Table 3 using the Average of Relative Citation (ARC), a measure that reflects citation rates and is normalised to account for differences among scientific specialities in the propensity to use references and receive citations. These data present the relative citation rate of OA publications overall, Gold OA and hybrid OA forms relative to publications in each discipline. A score above 1 denotes that papers are more cited than in the field overall, while a score below 1 means that these publications are less frequently cited. For instance, papers in Agriculture, Fisheries & Forestry receive roughly the same level of citation (0.98) in OA overall than they do usually (the base measure is 1.0 for whole set of papers in a discipline). Importantly though, Gold OA papers are cited only half as frequently on average (0.49), although self-archived and other hybrid forms are cited 28% more frequently than the discipline's average (1.28).

Table 3. Scientific impact (ARC) of OA publications, 2008

Field	All Publications	Green & Hybrid	Gold	OA
Agriculture, Fisheries & Forestry	1.00	1.28	0.49	0.98
Biology	1.00	1.35	0.55	1.15
Biomedical Research	1.00	1.17	0.84	1.13
Built Environment & Design	1.00	1.07	0.25	0.91
Chemistry	1.00	1.18	0.38	0.95
Clinical Medicine	1.00	1.66	0.59	1.40
Communication & Textual Studies	1.00	1.23	1.55	1.33
Earth & Environmental Sciences	1.00	1.04	1.19	1.05
Economics & Business	1.00	1.39	0.07	1.28
Enabling & Strategic Technologies	1.00	1.37	0.64	1.23
Engineering	1.00	1.49	0.13	1.41
General Arts, Humanities & Social Sciences	1.00	1.28	0.00	1.25
General Science & Technology	1.00	2.60	0.40	1.64
Historical Studies	1.00	1.10	0.22	0.84
Information & Communication Technologies	1.00	1.50	0.73	1.40
Mathematics & Statistics	1.00	1.11	0.71	1.07
Philosophy & Theology	1.00	1.28	0.61	1.18
Physics & Astronomy	1.00	1.21	1.05	1.19
Psychology & Cognitive Sciences	1.00	1.21	0.86	1.18
Public Health & Health Services	1.00	1.31	0.68	1.14
Social Sciences	1.00	1.38	0.52	1.18
Visual & Performing Arts	1.00	1.15	n.c.	1.02
All Publications	**1.00**	**1.36**	**0.59**	**1.21**

An overall OA advantage occurs in all but four disciplines (Agriculture, Fisheries & Forestry; Chemistry; Built Environment & Design; Historical Studies). Gold OA only presents a citation advantage in three disciplines (Communication & Textual Studies; Earth & Environmental Sciences; Physics & Astronomy), and in those disciplines, except for one (Physics & Astronomy), the citation advantage is greater in Gold OA than in hybrid OA forms. Hybrid OA forms always present a citation advantage.

These data require careful interpretation. First, many Gold journals are younger and smaller, and these factors have an adverse effect on the citation rate and the ARC. Authors frequently prefer reading and citing more established journals, and it is a difficult endeavour to start a journal from scratch. It takes time to build a reputation and to attract established authors. It is possible though that Gold journals might provide an avenue for less mainstream, more revolutionary science. If so, the signature would be a much greater level of variation between the more highly cited papers and the baseline with no citation. Also, the ARC is not scale-invariant, and larger journals have an advantage as this measure is not corrected sufficiently for journal size (namely, it is not a scale-independent measure). So it might not always be the Gold nature of journals that lowers their 'citedness'; instead several structural aspects might be at play. Even so, the Gold journal industry is young, and it is still difficult to separate the wheat from the chaff. In this respect, it might be useful for authors to examine Beall's List of 'potential, possible, or probable predatory scholarly open-access publishers' to lower one's risk of spending money on journals that do not espouse scientific publishing best practices.[168]

A last aspect of the analysis based on the pilot study data is the examination of OA availability per country (for EU27, EFTA, Accession countries, ERA, and four comparables). Please note that fractional counting was used here as it was deemed as potentially providing a more precise portrait of the situation. In fractional counting, if two authors are from separate countries, each country is given half a publication. In contrast, full paper counting would have ascribed one paper to each country. One advantage of fractional counting is that one can add the fractions for all countries' output in a table and obtain a total. A drawback is that statistics might not seem as intuitive. In the table, the fractions of papers are presented only for scores below 10 (for example, 11 papers; 3.2 papers). The EU27, EFTA, and ERA all have roughly the same level of OA as that observed at the world level, though there are noticeable differences among countries.

Excluding countries with less than 50 papers (sum of all the fractions), the EU countries with the greatest OA proportions are the Netherlands, Finland, Romania, Portugal, and the United Kingdom. The countries with the lowest rate of OA adoption are Hungary, the Czech Republic, Poland, Germany, and Denmark. In countries outside the EU27, it is noteworthy that the US seems to have passed the tipping point (50%). Even more salient is the proportion of 62%

[168] http://scholarlyoa.com/publishers/.

observed in Brazil. This is no doubt due largely to the exemplary work performed by Scielo, which plays a key role in the Southern hemisphere in making scientific knowledge more widely available.

Discussion

One has to be careful when interpreting the results presented in this paper as the methodological instruments are not fully developed, and results could vary with growing accuracy. As a general rule, further development of the 'trawler' will increase recall and therefore, the proportion of OA presented here will surely increase. Sample size can be fine-tuned to obtain a satisfactory level of *statistical precision* as the margins of error presented above were certainly high in several areas. Future exercise will balance the sample more carefully to augment the number of papers from the smaller countries and the presence of papers from the smaller disciplines. We also endeavour to develop a robust method to distinguish more clearly between Gold OA, Hybrid OA and non-fully Gold journals, and self-archiving ('Green OA'). This presents many challenges, and statistics should be presented on the condition that they must not be too inaccurate. Other authors have presented results suggesting that OA availability was only half as high as carefully and prudently measured here, and this is certainly a reminder that it might be preferable to be reflective. Previous authors have measured what was in databases, or what search engines were able to do. Our goal here is to estimate the proportion of peer-reviewed academic-level literature which is available for free. Measuring how well Google Scholar fares at identifying a part of this is certainly an interesting exercise in itself, but it does not address our central question.

Finding that the tipping point has been reached in open access is certainly an important discovery. This means that the publishing industry is undergoing revolutionary change and at a pace much faster than anticipated, in large part because previous measures of OA availability proved to be misleading. This means that aggressive publishers such as Springer are likely to gain a lot in the redesigned landscape, whereas those attached to the old days are likely to suffer and to lose market share. The impression gained in carrying out this study and developing our OA 'trawler' is that the tool plaza is being moved to the beginning of the publishing process, away from the back-end of the process, and thus from the libraries and closer to researchers. Despite what several authors thought, and argued for, green OA only appears to move slowly, whereas Gold OA and hybrid toll before the process as opposed to toll after are in the fast lane. Efforts need to be made to characterise these changes.

If the toll plaza changes from the end of the process to the front-end, one category of workers is likely to be highly affected: the university and research centre librarian. Librarians have been highly affected already by the shift from paper to digital media and losing the responsibility of spending the large sum paid in journal subscriptions will certainly create another large dent in their traditional sphere of responsibilities. If the tool plaza is just moved, it means that researchers will have control over the toll.

The market power will shift tremendously from the tens of thousands of buyers that publishers' sales staff nurtured to the millions of researchers that will now make the atomistic decision of how best to spend their publication budget. Much has been said about the cost of publishing in gold and hybrid OA, but one has to place this in perspective. The cost of academic papers in the US is about $125,000 on average (HERD divided by number of papers by academia) so adding or including a $2,000 publication fee in this envelope is certainly going to break the bank. The question is rather whether the switch to a more atomistic market will reduce, augment or leave unchanged the negotiating power of publishers. One will have to stay tuned and watch the gales of creative destruction at play.

Table 4. Proportion of OA availability by country, 2008

Group	Country	Papers	Green & Hybrid		Gold		OA	
			Papers	%	Papers	%	Papers	%
EU27	Austria	107	32	30%	10	10%	42	40%
	Belgium	185	77	42%	5.0	3%	82	44%
	Bulgaria	24	8.9	37%	2.0	8%	11	45%
	Cyprus	5.1	1.1	21%	0.9	18%	2.0	39%
	Czech Republic	92	28	30%	5.7	6%	33	36%
	Denmark	145	48	33%	5.2	4%	54	37%
	Estonia	15	3.6	25%	2.9	20%	6.5	45%
	Finland	96	41	43%	5.5	6%	47	49%
	France	773	254	33%	41	5%	295	38%
	Germany	1,016	316	31%	59	6%	375	37%
	Greece	143	50	35%	11	8%	61	43%
	Hungary	74	21	28%	2.8	4%	24	32%
	Ireland	63	23	36%	3.7	6%	26	42%
	Italy	604	214	35%	32	5%	246	41%
	Latvia	6.9	4.5	65%		0%	4.5	65%
	Lithuania	21	9.2	44%	3.0	14%	12	58%
	Luxembourg	1.4	0.1	4%	1.0	72%	1.1	76%
	Malta	2.5	1.1	45%	0.4	15%	1.5	60%
	Netherlands	313	150	48%	14	4%	164	53%
	Poland	229	56	25%	27	12%	83	36%
	Portugal	82	31	37%	8.2	10%	39	47%
	Romania	64	25	39%	5.8	9%	31	48%
	Slovakia	43	14	32%	7.0	16%	21	49%
	Slovenia	34	10	29%	3.8	11%	14	40%
	Spain	549	162	30%	55	10%	217	40%
	Sweden	220	73	33%	11	5%	84	38%
	United Kingdom	1,147	465	41%	59	5%	523	46%
	Total EU27	**6,055**	**2,118**	**35%**	**383**	**6%**	**2,500**	**41%**
EFTA	Iceland	8	3	35%	1	13%	4	48%
	Liechtenstein	1	1	100%		0%	1	100%
	Norway	95	31	32%	9	10%	40	42%
	Switzerland	194	65	34%	14	7%	79	41%
	Total EFTA	**296**	**99**	**33%**	**25**	**8%**	**124**	**42%**
Candidate	Turkey	327	65	20%	50	15%	115	35%
	Croatia	38	16	41%	4	11%	20	52%
	Macedonia	3	1	42%	2	58%	3	100%
	Total Candidate	**368**	**82**	**22%**	**56**	**15%**	**138**	**37%**
	Israel	137	59	43%	4	3%	63	46%
Total ERA		**6,855**	**2,358**	**34%**	**467**	**7%**	**2,825**	**41%**
Others	United States	4,524	2,140	47%	220	5%	2,360	52%
	Japan	1,072	349	33%	76	7%	425	40%
	Canada	598	243	41%	29	5%	273	46%
	Brazil	450	89	20%	212	47%	301	67%

Source:Computed by Science-Metrix

Acknowledgments

This research has received support from the European Commission.

References

Abad-García, M. F., Melero, R., Abadal, E., & González-Teruel, A. (2010). Self-archiving of biomedical papers in open access repositories. *Autoarchivo de artículos biomédicos en repositorios de acceso abierto, 50*(7), 431-440. doi: 10.1371/journal.pbio.0040157.

Antelman, K. (2004). Do open-access articles have a greater citation impact? *College & Research Libraries*, 65(5), 372–382.

Arlitsch, K., & O'Brien, P. S. (2012). Invisible institutional repositories: Addressing the low indexing ratios of IRs in Google Scholar. *Library Hi Tech*, *30*(1), 60-81. doi: 10.1108/07378831211213210.

Bailey Jr, C. W. (2005). The role of reference librarians in institutional repositories. *Reference Services Review, 33*(3), 259-267.

Bilder, G. (2003). Ingenta's economic and technical models for providing institutional OA archives. *Information Services and Use, 23*(2-3), 111-112.

Björk, B. C., Welling, P., Laakso, M., Majlender, P., Hedlund, T., & Gudnason, G. (2010). Open Access To The Scientific Journal Literature: Situation 2009. PLoS ONE, 5(6). doi: 10.1371/journal.pone.0011273.

Carbone, P. (2007). Consortium negotiations with publishers - Past and Future. *LIBER Quarterly*, 17(2).

Chan, D. L. H., Kwok, C. S. Y., & Yip, S. K. F. (2005). Changing roles of reference librarians: The case of the HKUST Institutional Repository. *Reference Services Review, 33*(3), 268-282. doi: 10.1108/00907320510611302

Chan, L., Devakos, R., & Mircea, G. (2005). *Workshop 2: Implementing and filling institutional repositories introduction*, Leuven-Heverlee.

Charles, L., & Booth, H. A. (2011). An Overview of Open Access in the Fields of Business and Management. *Journal of Business and Finance Librarianship, 16*(2), 108-124. doi: 10.1080/08963568.2011.554786

Craig, I. D., Plume, A. M., McVeigh, M. E., Pringle, J., & Amin, M. (2007). Do open access articles have greater citation impact? A critical review of the literature. *Journal of Informetrics, 1*(3), 239-248. doi: 10.1016/j.joi.2007.04.001.

Dubini, P., Galimberti, P., & Micheli, M. R. (2010). Authors publication strategies in scholarly publishing. In *ELPUB 2010 International Conference on Electronic Publishing*, Helsinki (Iceland), 16-18 June 2010.

Gargouri, Y., Larivière, V. Gingras, Y. and Harnad, S. (2012). Green and Gold Open Access Percentages and Growth, by Discipline. In Archambault, É, Gingras, Y. and Larivière. V. (2012). *Proceedings of 17th International Conference on Science and Technology Indicators*, Montréal: Science-Metrix and OST.

Gentil-Beccot, A., Mele, S., & Brooks, T. C. (2010). Citing and reading behaviours in high-energy physics. *Scientometrics, 84*(2), 345-355. doi: 10.1007/s11192-009-0111-1

Harnad, S. (2003). The research-impact cycle. *Information Services and Use, 23*(2-3), 139-142.

Harnad, S. (2008). Waking OA's "slumbering giant": The university's mandate to mandate open access. *New Review of Information Networking, 14*(1), 51-68. doi: 10.1080/13614570903001322.

Harnad, S. (2012). Open access: A green light for archiving. *Nature, 487*(7407), 302. doi: 10.1038/487302b

Harnad, S., & Brody, T. (2004). Comparing the impact of open access (OA) vs. non-OA articles in the same journals. *D-Lib Magazine*, 10(6).

Henderson, I. (2013). Open-Access and Institutional Repositories in Fire Literature. *Fire Technology, 49*(1), 155-161. doi: 10.1007/s10694-010-0198-1

Houghton, J. W. (2010). Economic implications of alternative publishing models: Self-archiving and repositories. *LIBER Quarterly, 19*(3-4), 275-292.

Hu, C., Zhang, Y., & Chen, G. (2010). Exploring a New Model for Preprint Server: A Case Study of CSPO. *Journal of Academic Librarianship, 36*(3), 257-262. doi: 10.1016/j.acalib.2010.03.010

Kurek, K., Geurts, P. A. Th M., & Roosendaal, H. E. (2006). The split between availability and selection: Business models for scientific information, and the scientific process? *Information Services and Use, 26*(4), 271-282.

Lakshmi Poorna, R., Mymoon, M., & Hariharan, A. (2012). A study of select open access journals and their business models listed in DOAJ in the fields of civil and structural engineering. *Journal of Structural Engineering (India), 39*(4), 458-468. doi: 10.1371/ journal.pone.0020961.

Larivière, V., Archambault, É., Gingras, Y. & Vignola-Gagné, É. (2006). The Place of Serials in Referencing Practices: Comparing Natural Sciences and Engineering with Social Sciences and Humanities, *Journal of the American Society for Information Science and Technology*, 57(8), 997-1004. doi: 10.1002/asi.20349.

Miguel, S., Bongiovani, P. C., Gómez, N. D., & Bueno-de-la-Fuente, G. (2013). Prospect for Development of Open Access in Argentina. *Journal of Academic Librarianship, 39*(1), 1-2. doi: 10.1016/j.acalib.2012.10.002

Morris, S. (2003). Open Publishing: How publishers are reacting. *Information Services and Use, 23*(2-3), 99-101.

Nyambi, E., & Maynard, S. (2012). An investigation of institutional repositories in state universities in Zimbabwe. *Information Development, 28*(1), 55-67. doi: 10.1177/0266666911425264

Pelizzari, E. (2004). Academic authors and open archives: A survey in the social science field. *Libri, 54*(2), 113-122.

Prosser, D. (2003). Institutional repositories and Open access: The future of scholarly communication. *Information Services and Use, 23*(2-3), 167-170.

Regazzi, John. (2004). The Shifting Sands of Open Access Publishing, a Publisher's View. Serials Review, 30(4), 275-280. doi: http://dx.doi.org/10.1016/j.serrev.2004.09.010
Repanovici, A. (2012). Professional profile of digital repository manager. *Library Hi Tech News, 29*(10), 13-20. doi: 10.1108/07419051211294473
Rowland, F. et al. (2004). Delivery, management and access model for e-prints and open access journals. *Serials Review, 30*(4), 298-303. doi: 10.1016/j.serrev.2004.09.006
Sawant, S. (2012). Past and Present Scenario of Open Access Movement in India. *Journal of Academic Librarianship*. doi: 10.1016/j.acalib.2012.11.007
Swan, A. et al. (2005). Developing a model for e-prints and open access journal content in UK further and higher education. *Learned Publishing, 18*(1), 25-40. doi: 10.1087/0953151052801479
Swan, A., & Brown, S. (2004). Authors and open access publishing. *Learned Publishing, 17*(3), 219-224. doi: 10.1087/095315104323159649
Woutersen-Windhouwer, S. (2012). The Future of Open Access Publishing in the Netherlands: Constant Dripping Wears Away the Stone. *Journal of Academic Librarianship*. doi: 10.1016/j.acalib.2012.11.015

TO WHAT EXTENT CAN RESEARCHERS' INTERNATIONAL MOVEMENT BE GRASPED FROM PUBLISHED DATA SOURCES?

Yasuhiro Yamashita[1] and Daisuke Yoshinaga[2]

[1] *yaya@jm.kj.yamagata-u.ac.jp,* [2] *yoshinaga@kdw.kj.yamagata-u.ac.jp*
Planning and Research Support Department, Yamagata University, 1-4-12, Kojirakawa-machi, Yamagata-shi, 99-8560 Yamagata (Japan)

Abstract
Using CVs or Short Bios in published resources, such as the Internet enables us to analyze many issues concerning researchers' careers. However, relatively little effort has been devoted to this area, and availability of this method, concerning target researchers, such as the sector of their institution, country of residence, their visibility (i.e., the impact of their publications) was not known enough. To trace this activity, we examine how many contributions that we were able to unveil in terms of authors' countries of origin by using two types of samples: highly cited papers and papers that have yet to be cited at all. Then, we analyze the influence of these researchers' international movement. The results show the full landscape of the movement's influence on national publication, the characteristics of each country in term of researchers' countries of origin, their research experience and acquisition of the research funds of both internationally moved and domestic researchers. Finally, we assess the limitations of the method and topic to be addressed concerning this method.

Conference Topic
Old and New Data Sources for Scientometric Studies: Coverage, Accuracy and Reliability (Topic 2).

Introduction
While the international movement of researchers is one of the most important issues of science and technology policies in many countries, almost no tool which can be used for comprehensive study has been utilized to analyze it. Thus, there have been relatively few studies devoted to quantitative studies of researcher's international movement in the area of scientometrics. Although researchers' identifier in conformity to ORCID, such as the "Researchers ID", might partially solve this problem partially in the future, their coverage is relatively low at present, and it does not cover their life history before their first publication. We explore to what extent we can obtain researchers' origin from published data source such as the Internet, then analyze the effect of movement of two kinds of researchers; (1) researchers who wrote highly cited papers in the period of 2004 to 2006; (2) those who wrote papers not cited in the same period.

There are several methods proposed for grasping researchers' history other than questionnaires (e.g. Cruz-Castro & Sanz-Menendez, 2010) and above-mentioned researchers' identifiers. The first method is using CVs or Short Bios published (e.g. Jonkers, 2010; Jonkers & Tijssen, 2008; Lepori & Probst, 2009). This method has advantages of nature not strain on surveyed researchers. To pursue researchers' careers more exhaustively, Dietz and his colleagues utilized the method of asking researchers to send their CVs via E-mail, using the method of collecting researchers' histories via the Internet concurrently (e.g. Dietz et. Al., 2000; Corley, Bozeman & Gaughan, 2003; Dietz & Bozeman, 2005). Although this method has the advantage of enabling collection with "unpublished' CVs, put away non-invasive features of collecting published CVs. Before using this method regularly, other non-invasive method should be assessed to minimize surveyed researchers' workloads.

As for other methods of collecting researchers' histories directly, extracting information on researchers' careers from journals which contain researchers' Short Bios, was utilized by few researchers such as Furukawa et al. (2011) and Yamashita et al. (2007). This method enabled collecting career information on researchers who might not tend to publish their information, such as employees of industry, Ph.D students or postdoctoral fellows. On the other hand, the area of research would be limited because journals which contain researchers Short Bios are rare.

Thomson Reuters' database, called "HighlyCited.com" also enabled grasping outstanding researchers' career information systematically (such as Ioannidis, 2004). While systematic data is attractive to analyze, this database cannot be used to pursue "ordinary researchers". So, alternative data sources are needed to grasp whole tendencies of a field/organization.

Alternative method for pursuing researchers was utilized by Laudel (2003). She insisted that using bibliographic database is more favorable than collecting CVs or questionnaires, since it made it possible to avoid influence of incompleteness of the CV data. This method seems to be effective for outstanding researchers or that with names not so common. However, as for researchers of Asian origin, their initials are frequently so common to identify individuals, and old records without direct linkage of each author to his/her affiliation in the Web of Science prevent us from accurate identification of each researcher, and to guarantee precision of data, each paper of respective researchers in the study should be consulted. Moreover, researchers can be pursued only after their first publications, therefore, it is difficult to grasp researchers' origin by this method. Thus, the bibliometrics method can be applied to a limited number of outstanding researchers.

On the other hand, method utilizing surname as an indicator of researchers' origin is also used to grasp researchers' origin (such as Jonkers, 2010; Lewison & Kundra, 2008). This method is suitable to utilize large samples since it does not need to seek each researcher's life history directly. However, surnames do not always indicate their own origin, but family origin since they do not contain any

information on the period of their migration. Therefore, this method can be used for pursuing researchers' family origin using large samples containing a certain error margin.

As above stated, each existing method to pursue researchers' movement or origin has its advantage and shortcomings, and is not used broadly because of its laborious nature and lack of coverage. Under the above-mentioned constraint, methods utilizing published information enables relatively free research design without constraint of selection of target field or researchers surveyed, without any labor of researchers surveyed. So, this method seems to be favorable for monitoring the situations of a certain research field. Thus, in this research, we validate this method by applying it to a basic engineering research field of artificial intelligence, which has the ability of broad industrial applications, can be expected to have researchers of broad country/sector origin, and in which relatively many publications are issued each year. For validating to what extent we can grasp the tendency of the whole field taking into consideration various researchers, we (1) seek all researchers' career information of sample papers and (2) explore both highly cited and uncited papers to examine how researchers' "visibility" affects their traceability. While "random sampling" from all publications in the field seems more valid to test the tendency of the entire field, we use two extreme samples for comparison, since it is difficult to obtain enough samples representing the field due to the high labor requirement.

Data and Methodology

We used publication data from the field of "Computer Science, Artificial Intelligence" published between 2004 and 2006 to make sure enough time had passed after publication, extracted from Web of Science database provided by Thomson Reuters. The retrieval was executed in January 2011. All top 1% cited papers (hereafter "highly cited papers") and randomly sampled papers without any citation (hereafter "uncited papers") were extracted from the whole sample. Since we had already presented about feature of highly cited papers concerning researchers' movement (Yamashita, 2011), we focus on the comparison between the two samples based on the previous study. Document type "Article" was used for analysis. Number of sample uncited paper 1% of whole sample almost as many as highly cited papers. CV or Short Bio of each author of sample papers was retrieved on the Internet or extracted from journals.

Of various information contained in each CV or Short Bio, we mainly focused on the origin of researchers, which can be extracted commonly.

Each author's contribution to each paper was counted fractionally, one by number of authors, to avoid overrating, mainly because we aimed to explore its applicability to assessment of national/institutional publication. Number of publications was counted focusing two periods; (1) number of papers of country "A" by their affiliating country (hereafter Ni(A)), and (2) number of papers of country "A" by researchers' origin (hereafter No(A)). To analyze effect of researchers' flow between two countries, (3) number of papers published in

country A by researchers originated in country B (hereafter Nc(A,B)) were also counted. Researchers' countries of origin were defined as that in which they obtained bachelor or equivalent degree or where they were born, along with the authors' previous study (Yamashita et al., 2007). As for European continental countries, in the case that countries in which researchers' obtained their master degrees were designated without bachelor, it was also accounted as their origin.
Identification rate was assessed by rate of papers of which researchers' origin was unveiled, since we focused on international movement of researchers. Thus, other information contained in CVs or Short Bios was excluded from our assessment of our identification rate.

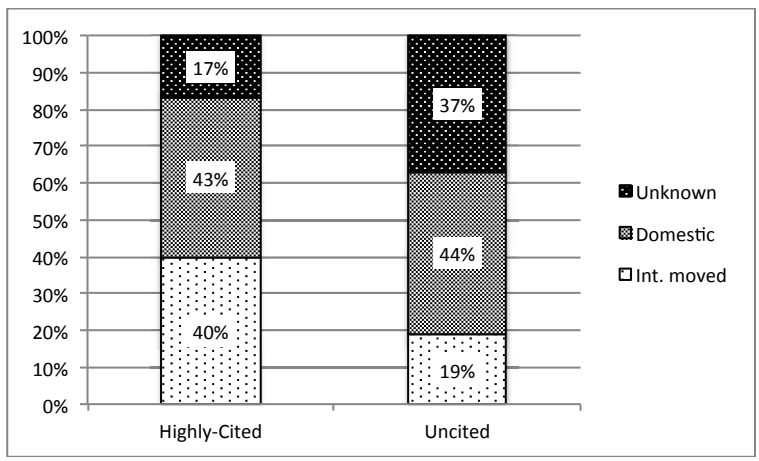


Figure 1. Breakdown of both highly cited and uncited papers by authors' origin.

Result

Identification rate of researchers' origin

Authors' origin was unveiled as much as 83.2% of the total 140 highly cited papers, while it was as low as 62.3% in 138 uncited papers (Figure 1). As for the breakdown of each sample, the highly cited papers consist of almost the same contribution of both international-moved researchers and domestic researchers (researchers who worked in their countries of origin), besides the uncited papers contain only 18% contributions of internationally-moved researchers. However, the contribution of unveiled researchers in uncited papers was as high as 37.7%, therefore it is difficult to estimate the contribution of internationally-moved researchers in uncited paper, which was much lower than that of highly cited papers.

Both the percentages of highly cited and uncited papers counted by authors' affiliation (Ni(A)) are presented in Figures 2 and 3 respectively. The top five countries of highly cited papers were, the US (sharing 44.8%), China (9.4%), the UK (6.0%) , Taiwan (5.0%) and France (4.7%) . Out of these five countries, the

share of the US has an oligopolistic share, however, more than half of these US papers were published by foreign-origin researchers. On the other hand, uncited papers were produced by researchers affiliated with institutes in Russia (13.3%), the US (10.4%), Canada (7.1%), Italy (6.9%) and China (6.8%) respectively. The distribution of uncited papers' countries was more equivalent without occupation by any specific country. The authors' origins were unveiled in most countries for highly cited papers, whereas they were not unveiled in many countries, such as Russia, China and Taiwan for uncited papers. The low coverage of Russia seemed to be caused by sparse information disclosure, not by the visibility of researchers, so the systematic bias might have occurred due to information inequality across countries.

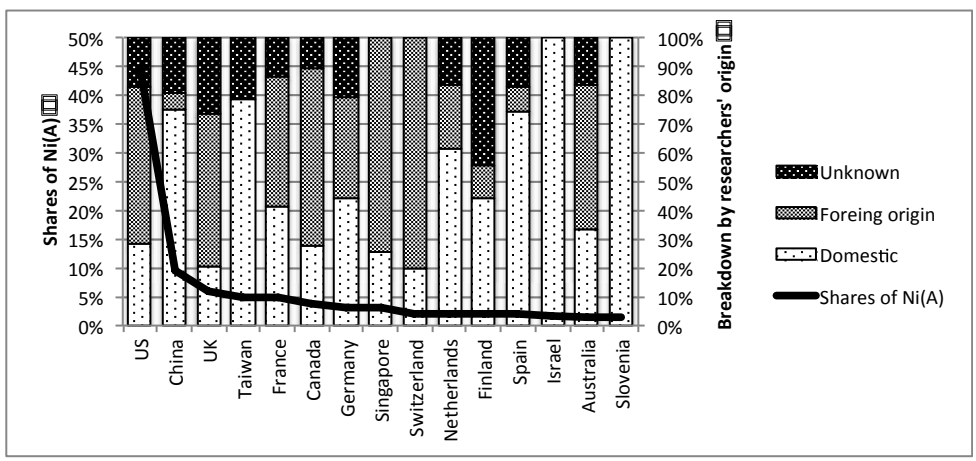


Figure 2. Shares of Ni(A) of highly cited publications and their breakdown by researchers' origin

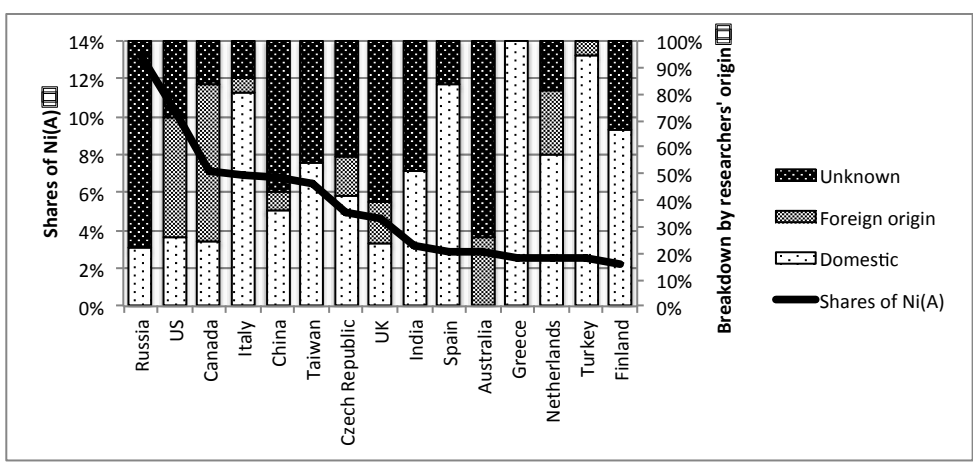


Figure 3. Shares of Ni(A) of uncited publications and their breakdown by researchers' origin

How does coverage should change by impact of papers? A scatter plot of countries which appeared in both highly cited and uncited papers showed the change of coverage according to the impact of the papers (Figure 4). Canada, Spain and Netherlands did not change their coverage by citation impact, and coverage of both highly cited and uncited papers exceeded 80%. Countries of which coverage changed significantly were Taiwan, China and UK. Their coverage of researchers' origin remained at approximately 40%. Thus, advantaged researchers should be selected if their origin or movement is analyzed in this field.

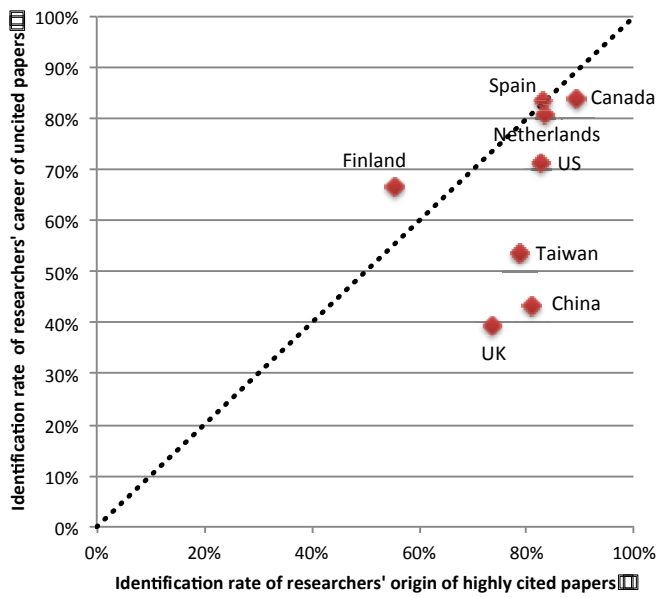


Figure 4. Identification rate of both highly cited and uncited papers.

In addition, it seemed reasonable to assume that researchers' disclosure of their origin should depend on sectors of their institutes, such as, university, public research institute, industry, and so on. For example, university researchers seem to have a tendency to disclose their CVs on their institutes' websites.

Figure 5 shows the identification rate of researchers' origin by the three most productive sectors (university, public research institute, and industry). The identification rate was increased by citation impacts in all three sectors. The largest difference of the three sectors appeared in the public research institute, which was mainly caused by the Russian researchers of this sector who had a tendency not to disclose their career information. If all Russian researchers were removed from the uncited papers, the identification rates of the university, public research institute, and industry were relatively similar (70%, 65% and 56% respectively). The identification rate of both highly cited and uncited papers of

industrial researchers was close to each other: they tend not to expose individual researchers, regardless of their contributions

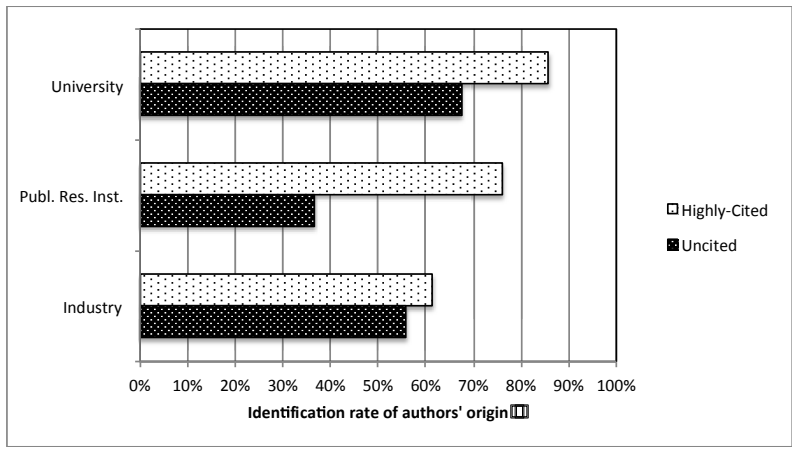


Figure 5. Identification rate of researchers' origin by their institutional sectors.

International movement of authors of both higyly-cited and uncited papers

In this section, we report our comparison of the influence of researchers' international movement on the publication of highly cited papers in comparison that of uncited papers. The identification rate of uncited papers was relatively low, and there were countries of which identification rates were very low, so only limited interpretation was possible. However we attempt to grasp tendencies as far as possible, by analysis excluding data of which authors' origin was not unveiled.

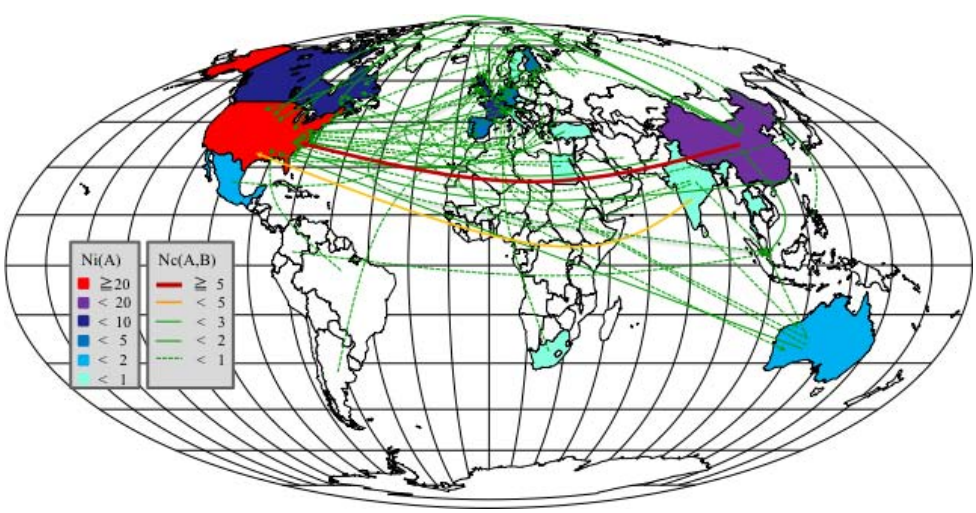


Figure 6. Influence of researchers' movement on national highly cited publications.

A glance at the overall influence of researchers' flow between two countries revealed a noticeable number of papers published by researchers who moved from China or India to the US (Figure 6). They are both giants in terms of sending massive amounts of research personnel to other countries, particularly the US and European countries. The difference between the two movements is that China produced its own highly cited publications (at a rate that is second only to that of the United States), whereas India produced limited ones.

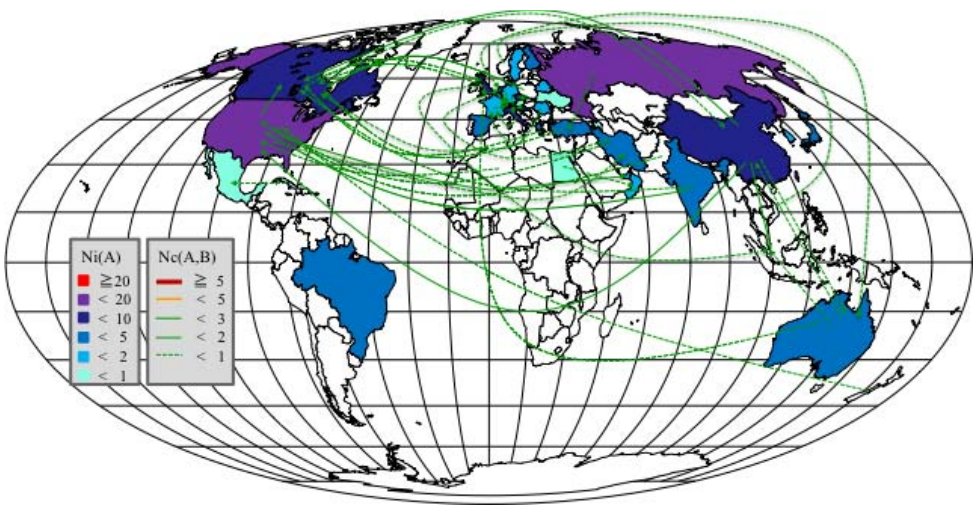


Figure 7. Influence of researchers' movement on national uncited publications.

On the other hand, as for uncited papers, there were relatively low numbers of publications by internationally moved researchers (Figure 7). Russia which did not publish any highly cited papers, showed the highest Ni(A), however, it had no publication by foreign originated researchers. In contrast, India which provided highly cited researchers to other countries, published uncited papers domestically. In the case of the Russian researchers in our sample, they published all of their articles in the "Journal of Computer and Systems Sciences International" which is the English-translated version of the Russian theoretical journal. Moreover, most of the articles indexed into the field of "Computer Science, Artificial Intelligence" in the WoS by Russian researchers were published in the journal. So it can be presumed to be one of the most major Russian journals in the field and the Russian original version might be cited by domestic journals which might not be indexed into the WoS. Therefore, it should be taken into account that there might be some unavoidable geographic bias of citation.

Figures 8 and 9 show the fact that most countries except for the US, provided researchers who published highly cited papers to other countries while uncited papers were published by domestic researchers who did not move to other countries than that of their origin. Comparison of highly cited and uncited papers revealed the fact that researchers who published highly cited tended to move to

other countries from their origin. The reason might be both that researchers with ability to produce high-impact tended to seek leading-edge research environment that utilize their ability, or that brilliant researchers had more chance to carry out their research in advanced countries with leading-edge research environment. Besides, two Asian research personnel providing giants, China and India, showed contrasting features; researchers of Chinese origin produced both highly cited and uncited papers domestically in certain percentages, while those of India published most of highly cited papers abroad and half of uncited papers domestically. Thus, it could be suggested that China facilitated more leading-edge research environment. Researchers of US origin produced all their highly cited papers domestically, therefore, it was suggested that the US attracted both domestic and foreign-origin outstanding researchers.

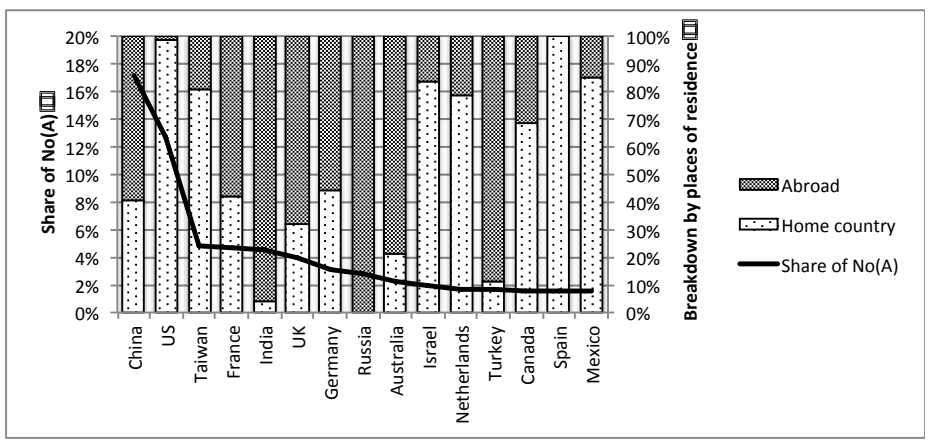


Figure 8. Shares of No(A) of highly cited publications and their breakdown by their places of researchers' residence

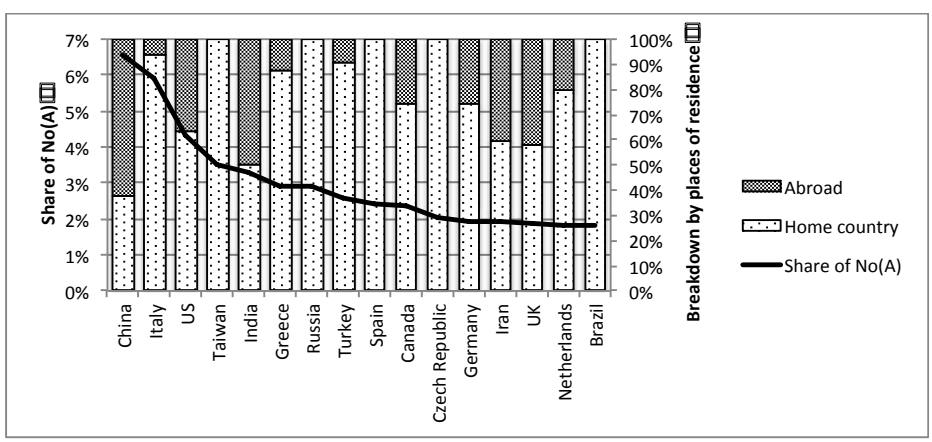


Figure 9. Shares of No(A) of uncited publications and their breakdown by their places of residence

What types of careers did the authors of highly cited papers have? There were several variables indicating researchers' career, such as current or former positions, experience of doing research abroad and number of years of active research. Here, we present only years of research experience, since we could not apply the former two indicators to our study because of our research design; many Short Bios did not reveal positions or accurate working periods, and two different data sources, CVs/Short Bios and the Web of Science database were used for our study, so it was difficult to learn of the researchers' experience before the publication of specific papers. The years of research experience was defined as number of years after researchers' obtained their bachelor or equivalent degrees, along with our previous study (Yamashita et al., 2007). Here, we used a head count not the researcher's number of publications.

The mean years of experience of authors of highly cited papers was 15.8 (15.1 years for internationally moved researchers, 16.2 years for domestic researchers), and was relatively shorter than that of uncited (18.6 years for all, 18.4 for internationally moved, and 18.7 for domestic). Distribution of experience years showed that researchers with 6 to 15 years of experience occupied half of internationally moved researchers who published highly cited papers, while those with same years of experience occupied only 34% of internationally moved researchers published uncited papers. Although, more detailed analysis of researchers position was needed for securing accuracy, the result suggested that many younger researchers before earning tenure published their highly cited papers abroad.

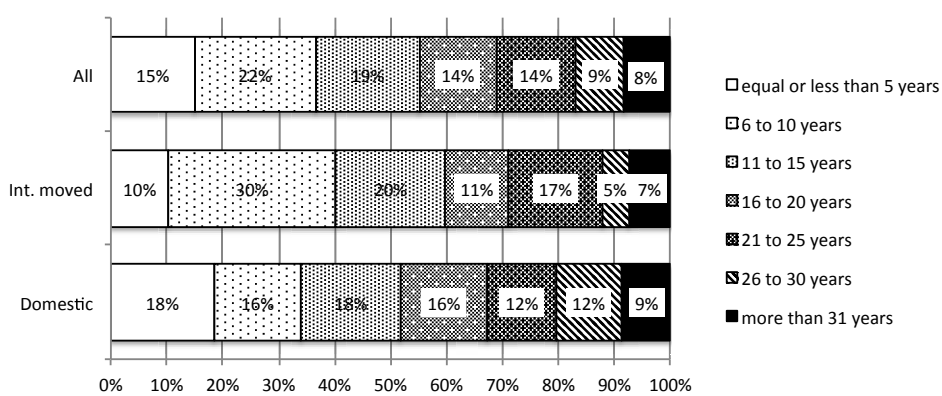


Figure 10. Years of research experience of authors of highly cited papers.

Developed countries are now recognizing excellent foreign-origin researchers as engines of knowledge production, which raised a question in our minds: is there any difference between the environments of researchers who published highly cited papers and those who published uncited papers? We analyzed research funds, which are among the most important resources for conducting research.

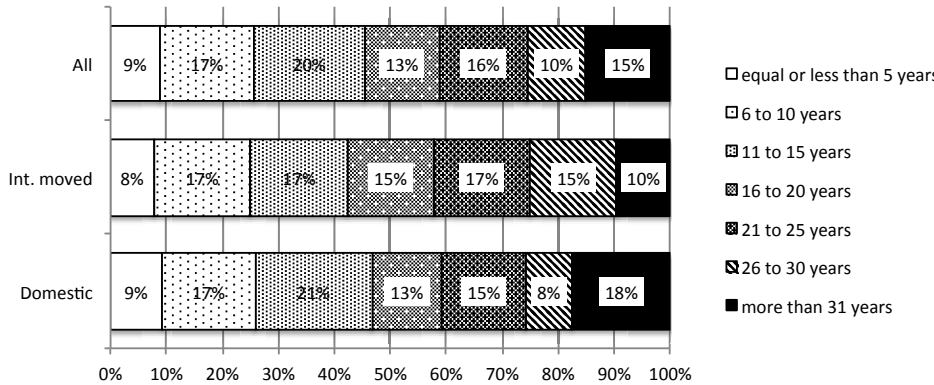


Figure 11. Years of research experience of authors of uncited papers.

We gathered funding information from the "Acknowledgements" section of each paper. For some studies, only a fraction of the authors had been supported by funding; yet in many of such cases, the funds were acknowledged on a general level. They were not linked to specific authors. So, we chose to attribute all funding information to every author of a paper. The authors indicate their acknowledgements voluntarily, so papers that did not designate any funds were not necessarily unfunded. Because of the abovementioned restrictions, only large differences between the two groups might be significant.

The rate of papers (excluding those that could not be obtained) with funding information were 45% for highly cited papers, which is 10% higher than that of uncited. It was difficult to identify the difference between the two groups, because the difference was not so large.

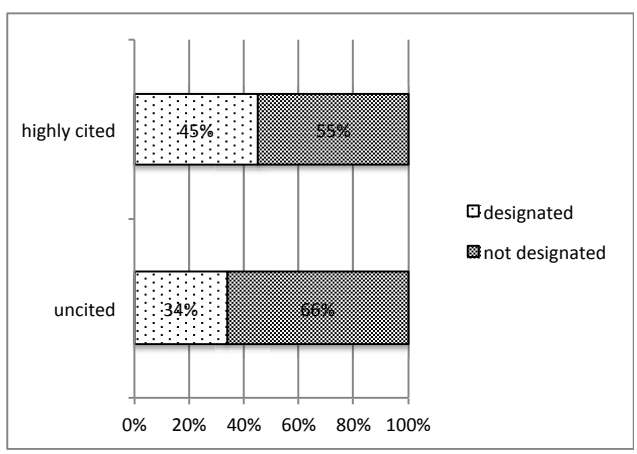


Figure 12. Rate of funding indication into the "Acknowledgements" section.

However, a breakdown according to status of movement showed a difference between them. For highly cited authors, there were almost no differences between the two groups, whereas there were tremendous differences between uncited authors. Thirty-six percent of domestic researchers were funded, and only 20% of international researchers were funded. Therefore, it seems that highly cited papers tended to be produced by funding that international researchers could obtain as easily as could domestic researchers, regardless of the amount.

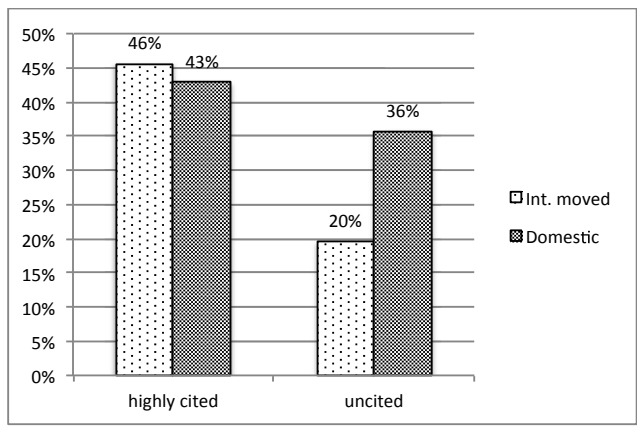


Figure 13. Rate of funding indication into the "Acknowledgements" section according to researchers' international movement status.

Conclusion

Our study was limited to a relatively small sample in a specific field, so our results might not apply to all fields. However, at least for highly cited papers, especially those written by university researchers, this method based on published researchers CVs or Short Bios proved to be effective from our study. However, there were countries of where identification rates were systematically low, such as Russia. Such countries might pose a strong bias to data, so in the case of random sampling from a whole population, analysts should be careful.

Our study also revealed the fact that industrial researchers do not tend to unveil their career information. So researchers' CVs or Short Bios should be collected periodically, especially for the fields relating to industrial application, because many researchers move across sectors.

On the other hand, our analysis of international movement revealed that both domestic and internationally moved researchers contribute to national publications almost equally. Especially, the influence of researchers' flow from the two Asian giants (China and India) to the US was observed. However, profiles of them were contrastive; Chinese researchers produced their highly cited papers domestically to some extent, whereas Indian researchers tended to produce most of highly cited papers in abroad. What caused their contrastive natures? One possible interpretation is their policies concerning the use of foreign outstanding

research personnel originated in them. Jonkers (2008b) reported that China implemented many programs to attract overseas Chinese scholars to return their home country whereas India had not been nearly as keen to do so. To analyze this, careers of researchers should be analyzed tolerating certain amount of error attributed to a lack of order or accurate period of them in researchers' CVs or Short Bios.

Our analysis of researchers' years of experience and of their funding for their papers revealed that highly cited papers were produced utilizing researchers with 6 to 15 years after taking bachelor degree, and produced in the environment in which foreign-origin researchers could use funds as much as domestic researchers either directly or indirectly. Although it is kept in mind that the research funds designated in papers were not necessarily distributed to all authors, and their amount was out of consideration, these results suggested that highly cited papers were produced in an environment where young excellent foreign-origin researchers were utilized and were allowed to take research funds regardless of their origin.

The present research utilized uncited papers for comparison with highly cited and could depict some natures of highly cited papers. However, factors which leave papers uncited seemed to be so diverse that assessment of them seemed problematic. As MacRoberts & MacRoberts (2010) pointed out, uncited papers were not necessary to be not used. While many papers published by researchers working in Russian institutes were not cited in our study, it should be taken into account that they are published in a journal translated from an original Russian journal. Thus if they were assessed, citation that the original Russian paper obtained should be taken into account.

Finally, the present study aimed to analyze the influence of human capital on national publications of highly cited papers, so we focused on output not on researchers. However, authors of highly cited papers do not always publish highly cited papers, and authors of uncited papers might have a chance to publish highly cited papers during their long career. Although we could not find any authors who published both highly cited and uncited papers in our sample, such situations might occur in other contexts Thus, our research design required each author to be dealt with as an aggregation, since its population contained inherent noise to some extent.

Quantitative analyses based on researchers' career information provide us with abundant suggestions, that cannot be obtained by bibliometrics solely. However, its laborious nature and low coverage would be caused by the nonstandardized formatting of CVs and Short Bios. Therefore, for analyzing researchers dealing as collective, it is desirable to develop a method for efficient data gathering and coding.

Acknowledgments

This research was supported by the JST/RISTEX research funding program titled "Science of Science, Technology and Innovation Policy" and by JSPS KAKENHI Grant Number 23500304.

References

Corley, E.A., Bozeman, B. & Gaughan, M. (2003). Evaluating the impacts of grants on women scientists' careers: The curriculum vita as a tool for research assessment. In P. Shapira and S. Kuhlmann (Eds), *Learning from science and technology policy evaluation: Experiences from the U.S. and Europe*, (pp.293-315).

Cruz-Castro, L. & Sanz-Menendez, L. (2010). Mobility versus job stability: Assessing tenure and productivity outcomes. *Research Policy,* 39, 27-38.

Dietz, J.S., Chompalov, I., Bozeman, B. & Park, J. (2000). Using the curriculum vita to study the career paths of scientists and engineers: An exploratory assessment. *Scientometrics*, 49, 419-442.

Dietz, J. S., & Bozeman, B. (2005). Academic careers, patents, and productivity: industry experience as scientific and technical human capital. *Research Policy*, 34, 349-367.

Furukawa, T., Shirakawa, N. & Okuwada, K. (2011). Quantitative analysis of collaborative and mobility networks. *Scientometrics* 87, 451-466.

Ioannidis, J.P.A., (2004). Global estimates of high-level brain drain and deficit. *FASEB Journal*, 18, 936-939.

Jonkers, K. (2008a). Chinese researchers returning home: Impacts of international mobility on research collaboration and scientific productivity, *Scientometrics*,77, 309-333.

Jonkers, K. (2008b). A comparative study of return migration policies targeting the highly skilled in four major sending countries, *Analytical Report*, *MIREM-AR 2008/05*, European University Institute.

Jonkers, K. (2010). *Mobility, Migration and the Chinese Scientific Research System*, Oxford: Routledge.

Lepori, B. & Probst, C. (2009). Using curricula vitae for mapping scientific fields: A small-scale experience for Swiss communication sciences. *Research Evaluation*, 18, 125-134.

Lewison, G. & Kundra, R.(2008). The internal migration of Indian scientists, 1981–2003, from an analysis of surnames. *Scientometrics*, 75, 21-35.

Laudel, G. (2003). Studying the brain drain: Can bibliometric methods help?, *Scientometrics* 57, 215-237.

MacRoberts, M.H. & MacRoberts, B.R. (2010), Problems of citation analysis: A study of uncited and seldom-cited influences. *Journal of the American Society for Information Science and Technology*, 61, 1-13.

Sandström, U. (2009). Combining curriculum vitae and bibliometric analysis: mobility, gender and research performance, *Research Evaluation*, 18, 135-142.

Yamashita, Y., Tomizawa, H., Ueno, S. & Kondo, M.(2007, June). Influence of the International migration of researchers on national publications in three fields of engineering. Poster session presented at *the 11th International Conference on Scientometrics and Informetrics*. Madrid.

Yamashita, Y. (2011, July). An attempt to grasp researchers' international migration. Poster session presented at *the 13th International Conference on Scientometrics and Informetrics*. Durban.

TO WHAT EXTENT IS THE H-INDEX INCONSISTENT? IS STRICT CONSISTENCY A REASONABLE REQUIREMENT FOR A SCIENTOMETRIC INDICATOR?

Yuxian Liu[1]

[1] *yxliu@tongji.edu.cn*
Tongji University, Tongji University Library, Siping Street 1239,200092 Shanghai, (China)

Abstract

We use a combinatorial approach to calculate the probability of inconsistency of the h-index and try to figure out the factors that influence the probability of the occurrence of this inconsistency. We observe that only when the number of new papers that two authors publish is not smaller than the difference of the h-indices of these two authors, and the numbers of citations that these papers have received are not smaller than the larger h-index of two authors, the inconsistency problem of the h-index can occur. We hence argue that inconsistency of the h-index is caused by the progress that the laggard is making so that the laggard can catch up to the trendsetter. In this sense we suggest that some inconsistency should be tolerated when we design a scientometric indicator to measure the development of science or the progress of a scientist. We show by our calculations that factors such as the h-indices of two authors, the difference between their h-indices, the maximum numbers of citations that two authors can receive, the number of new papers that the authors publish later and the number of citations received by these new papers all influence the probability of inconsistency of the h-index.

Introduction

The notion of an *indicator* is a powerful concept in scientometric research. *Validity* and *reproducibility* are two basic requirements to decide if an indicator is acceptable. Validity means that one has to make sure that one really measures what is intended to be measured. *Reproducibility* means that under identical conditions results must be identical.

Recently consistence and independence are proposed as requirements for the acceptance of a research indicator (Bouyssou&Marchant, 2010, 2011a, 2011b; Marchant, 2009a, 2009b, Waltman and van Eck, 2009a, 2009b, 2011). However, the meaning of (and difference between) the terms consistence and independence is not always clear. Here we use the term consistence with respect to ranking two authors, if the following rule holds.

> Consider two authors, A1 and A2, where A1 is considered strictly better than A2 according to the given ranking method for authors; if now the two authors improve their paper/citation record by the same (absolute)

amount, then A1 must stay strictly better than A2 (Bouyssou & Marchant, 2011b).

The notion of congruousness as defined by Rousseau is slightly different from consistence. Rousseau (2011) uses a reference set to determine scores and rankings. This reference set is subdivided into K disjoint classes; if a document belongs to class K, then it receives a score x_k. If the reference set does not change, then all class borders do not change and the score of an element will not change. In this case an indicator will be consistent.
However, if indicators are designed to measure the development of science, it is unavoidable to add one or more new articles that change the reference set by which we define classes. We then have to know what happens to the value of an indicator if a new articleis published. Rousseau (2011) argued that it was always possible—by exploiting these small changes—to prove inconsistency. Here we ask the questions: "To what extent do these small changes give rise to inconsistency? Which factors influence the extent of inconsistency?"
We use the h-index (see definition below) to illustrate the notion of inconsistency.

Questions

The h-index is defined as follows: A scientist has an *h*-index of *h* if *h* of his papers each have at least *h* citations and his remaining papers each have fewer than *h* + 1 citations (Hirsch 2005).
If we rank the papers of scientist S according to the number of citations each of these papers has received then the first h paper must have at least h citations. Thepapers ranked between rank 1 and rank h form the h-core. Papers ranked after rank h form the h-tail(Hirsch 2005, Liu & Rousseau 2009).

For our investigations we assume that, in a given field F, there exists a maximum number of citations over the period of interest. As an example we take this maximum equal to 7. In general this maximum is denoted as M_F. We assert that this assumption is reasonable since the maximum number of citations in any field is finite and can be obtained from the used database. Now we randomly select two authors, author A and author B, in this field and ask in how many ways numbers of citations of publications in the h-core can lead to the h-indices of these two authors. The h-index of author A is h_A , and the h-index of author B is h_B. As a case study we assume that the h-index of author A is 5 and that of author B is 3. There are no further restrictions so that the number of citations of an article written by A or B both author A and author B can reach the field's maximum number of citations.

Suppose that each of the above authors publishes the same number of additional papers and that, for each of these papers, each author receives exactly the same number of citations. How will the new papers change the h-index of these two authors? In which situations can the h-index of author A become higher than the

h-index of author B? In which situations can this not happen? To what extent is the h-index inconsistent? Is strict consistency a reasonable requirement for a scientometric indicator?

Arrays of numbers of citations of the papers in the h-core that can make the h-index of an author be equal to h

There are h papers in the h-core if the h-index of an author is h. The h numbers of received citations constitute an array. How many arrays can make the h-index of an author be equal to h? Concretely how many arrays can make the h-index of author A be h_A? How many arrays can make the h-index of author B be h_B?

This question can be formulated as a combinatorial problem. For author A, these h_A numbers of citations can be any h_A integers in the set of integers in the interval $[h_A, M_F]$. So $h_A$integers are repeatedly selected from the set of integers in the interval $[h_A, M_F]$ and then ranked in the h_A positions. In how many different ways can this be done?

We use $\#(h_A M_F)$ to denote the number of selections from the set of integers in the interval $[h_A, M_F]$.

$$\#(h_A M_F) = \binom{M_F - h_A + 1 + h_A - 1}{h_A} = \binom{M_F}{h_A} = \frac{M_F!}{(M_F - h_A)!\, h_A!}$$

This expression is sometimes referred to as an h_A-combination with repetitions.

For author A, $h_A = 5$; these 5 integers are repeatedly selected from the set {5,6,7}.

$$\#(h_A M_F) = \binom{7}{5} = 21$$

There 21 arrays are shown in table 1.

Table 1. Arrays of the numbers of citations of the papers in the h-core of the author A

7,7,7,7,7	7,7,6,6,6	7,7,7,7,5	7,5,5,5,5	6,6,6,5,5	7,7,7,6,5	7,7,6,5,5
7,7,7,7,6	7,6,6,6,6	7,7,7,5,5	5,5,5,5,5	6,6,5,5,5	7,7,6,6,5	7,6,6,5,5
7,7,7,6,6	6,6,6,6,6	7,7,5,5,5	6,6,6,6,5	6,5,5,5,5	7,6,6,6,5	7,6,5,5,5

In our study the numbers of citations of the papers in h-tail don't influence the change of the h-index, so we do not consider them.

For author B, $h_B = 3$ these 3 integers are repeatedly selected from the set of integers in the interval [3,7].

$$\#(h_B M_F) = \binom{7}{3} = 35$$

These 35 arrays are shown in table 2.

Table 2. Arrays of the numbers of citations of the papers in the h-core of author B

3,3,3	5,4,3	6,3,3	6,5,5	7,3,3	7,5,5	7,7,3
4,3,3	5,4,4	6,4,3	6,6,3	7,4,3	7,6,3	7,7,4
4,4,3	5,5,3	6,4,4	6,6,4	7,4,4	7,6,4	7,7,5
4,4,4	5,5,4	6,5,3	6,6,5	7,5,3	7,6,5	7,7,6
5,3,3	5,5,5	6,5,4	6,6,6	7,5,4	7,6,6	7,7,7

The number of the new papers and the numbers of citations of these new papers that can give rise to the problem of inconsistency of the h-index

If author A and author B publish the same number of new papers receiving the same numbers of citations, how many papers must author A and author B publish and how many citations must these papers receive so that the h-indices of these two authors will lead to an inconsistency problem?

We use #(P) to denote the number of new papers that two authors publish and we use h_A to denote the new h-index of author A and h_B to denote the new h-index of author B. Then we have the following inequalities:

$$h_A > h_B$$
$$h_A \leq \#(P) + h_A$$
$$h_B \leq \#(P) + h_B$$
$$h_A - h_B \geq h_A - (\#(P) + h_B)$$
$$\text{As } h_A \geq h_A$$
$$h_A - h_B \geq h_A - h_B - \#(P)$$
$$\text{If now} \#(P) < h_A - h_B$$
$$\text{then } h_A - h_B > 0.$$

So the problem of inconsistency will not occur in this case.

We use $\#(C_N)$ to denote the number of citations received by the two authors' new papers.If the number of citations received by a new paper is not larger than the smaller h-index of the two authors (assume it ish_B), $\#(C_N) \leq h_B$, then the h-indices of these two authors will not change. The problem of inconsistency cannot occur.

If the numbers of citations that these papers receive are larger than the smaller h-index but less than the larger h-index of the two authors,$h_B < \#(C_N) < h_A$, these papers cannot make the h-index of author B increase to h_A. This will also not give rise to an inconsistency.

These considerations lead to proposition A.

Proposition A: if the number of new papers that two authors publish is smaller than the difference of the h-indices of these two authors, this will not give rise to an inconsistency. If the numbers of the citations these papers have received are

less than the larger h-index of two authors, then again there will be no inconsistency.

Hence, only when the number of new papers that two authors publish is not smaller than the difference of the h-indices of these two authors and the numbers of the citations these papers have received are larger than or equal to the larger h-index of the two authors, the problem of inconsistency can occur.

We now only consider situations for which the inconsistency problem can occur and check to what extent the h-index is inconsistent, i.e. we only consider situations such as $\mathrm{h_A} \leq \#(\mathrm{C_N}) \leq \mathrm{M_F}$.

Change of the arrays of the numbers of citations of the papers in the original h-core and the numbers of citations of new papers

We now calculate the number of new arrays of each author formed by the numbers of citations of the papers in the h-core and the new papers the author publish. The corresponding numbers of citations of the additional papers of author A and author B are the same. We use $\#(\mathrm{C_1})$ to denote the number of citations the paper that ranked at the first place has received. And also $\mathrm{h_A} \leq \#(\mathrm{C_1}) \leq \mathrm{M_F}$. how many arrays can be formed by these numbers of citations of the papers in the h-core and the new paper? This is equivalent to the combinatorial problem of determining in how many different ways $\mathrm{h_A} + 1$ integers can be repeatedly selected from the set of integers in the interval $[\mathrm{h_A}, \mathrm{M_F}]$.

We use $\#(\mathrm{N1_A})$ to denote the number of combinations of these numbers of citations of the papers in the h-core of author A and of the one new paper by author A.

$$\#(\mathrm{N1_A}) = \binom{\mathrm{M_F} - \mathrm{h_A} + 1 + h_A + 1 - 1}{h_A + 1} = \binom{\mathrm{M_F} + 1}{h_A + 1} = \binom{8}{6} = 28$$

These 28 arrays are shown in table 3.

Table 3. Arrays of the numbers of citations of the papers in the h-core and one new paper of author A

7,7,7,7,7,5	7,6,6,6,6,5	7,7,5,5,5,5	6,6,6,5,5,5	7,7,6,6,5,5	7,6,5,5,5,5	7,7,6,6,6,6
7,7,7,7,6,5	6,6,6,6,6,5	7,5,5,5,5,5	6,6,5,5,5,5	7,6,6,6,5,5	7,7,7,7,7,6	7,6,6,6,6,6
7,7,7,6,6,5	7,7,7,7,5,5	5,5,5,5,5,5	6,5,5,5,5,5	7,7,6,5,5,5	7,7,7,7,6,6	6,6,6,6,6,6
7,7,6,6,6,5	7,7,7,5,5,5	6,6,6,6,5,5	7,7,7,6,5,5	7,6,6,5,5,5	7,7,7,6,6,6	7,7,7,7,7,7

We use $\#(\mathrm{N1_B})$ to denote the number of combinations of these numbers of citations of the papers in the h-core and one new paper of author B. This new paper also receives $\#(\mathrm{C_1})$ citations.Then how many combinations do these numbers have? This is equivalent to finding the number of ways in which we

select $h_B + 1$ integers repeatedly from the set of integers in the interval $[h_B, M_F]$ minus the number of ways we select $h_B + 1$ integers repeatedly from the set of integers in the interval $[h_B, h_A - 1]$.

$$\#(N1_B) = \binom{M_F - h_B + 1 + h_B + 1 - 1}{h_B + 1} - \binom{h_A - h_B + h_B + 1 - 1}{h_B + 1}$$
$$= \binom{M_F + 1}{h_B + 1} - \binom{h_A}{h_B + 1} = \binom{8}{4} - \binom{5}{4} = 65$$

These 65 arrays are shown in table 4.

Table 4. Arrays of the numbers of citations of the papers in the h-core and one new paper of author B

7,7,7,7	6,5,5,5	7,6,4,4	7,6,5,3	5,4,4,3
7,7,7,6	5,5,5,5	6,6,4,4	6,6,5,3	7,7,3,3
7,7,6,6	7,7,7,4	7,5,4,4	7,5,5,3	7,6,3,3
7,6,6,6	7,7,6,4	6,5,4,4	6,5,5,3	6,6,3,3
6,6,6,6	7,6,6,4	5,5,4,4	5,5,5,3	7,5,3,3
7,7,7,5	6,6,6,4	7,4,4,4	7,7,4,3	6,5,3,3
7,7,6,5	7,7,5,4	6,4,4,4	7,6,4,3	5,5,3,3
7,6,6,5	7,6,5,4	5,4,4,4	6,6,4,3	7,4,3,3
6,6,6,5	6,6,5,4	7,7,7,3	7,5,4,3	6,4,3,3
7,7,5,5	7,5,5,4	7,7,6,3	6,5,4,3	5,4,3,3
7,6,5,5	6,5,5,4	7,6,6,3	5,5,4,3	7,3,3,3
6,6,5,5	5,5,5,4	6,6,6,3	7,4,4,3	6,3,3,3
7,5,5,5	7,7,4,4	7,7,5,3	6,4,4,3	5,3,3,3

We now calculate the number of arrays formed by the numbers of citations in the h-core of each author and two new papers published by each author. The numbers of citations of two new papers of author A are the same as those received by author B. We use $\#(C_{21})$ to denote the number of citations received by one paper and use $\#(C_{22})$ to denote the number of citations the other paper has received. We have: $h_A \leq \#(C_{21}) \leq M_F$, $h_A \leq \#(C_{22}) \leq M_F$.

How many arrays can be formed by these numbers of citations of author A? This problem is equivalent to the combinatorial problem of finding in how many different ways $h_A + 2$ integers can be repeatedly selected from the set of integers in the interval $[h_A, M_F]$.

We use $\#(N2_A)$ to denote the number of combinations of these numbers of citations of the papers in the h-core and two new papers of author A.

$$\#(N2_A) = \binom{M_F - h_A + 1 + h_A + 2 - 1}{h_A + 2} = \binom{M_F + 2}{h_A + 2} = \binom{9}{7} = 36$$

These 36 arrays are shown in table 5.

Table 5. Arrays of the numbers of citations of the papers in the h-core and two new papers of author A

7,7,7,7,7,7,7	7,7,7,7,6,6,6	7,7,7,6,6,5,5	7,7,6,6,5,5,5	7,6,6,6,5,5,5	6,6,6,6,6,5,5
7,7,7,7,7,7,6	7,7,7,7,6,6,5	7,7,7,6,5,5,5	7,7,6,5,5,5,5	7,6,6,5,5,5,5	6,6,6,6,5,5,5
7,7,7,7,7,7,5	7,7,7,7,6,5,5	7,7,7,5,5,5,5	7,7,5,5,5,5,5	7,6,5,5,5,5,5	6,6,6,5,5,5,5
7,7,7,7,7,6,6	7,7,7,7,5,5,5	7,7,6,6,6,6,6	7,6,6,6,6,6,6	7,5,5,5,5,5,5	6,6,5,5,5,5,5
7,7,7,7,7,6,5	7,7,7,6,6,6,6	7,7,6,6,6,6,5	7,6,6,6,6,6,5	6,6,6,6,6,6,6	6,5,5,5,5,5,5
7,7,7,7,7,5,5	7,7,7,6,6,6,5	7,7,6,6,6,5,5	7,6,6,6,6,5,5	6,6,6,6,6,6,5	5,5,5,5,5,5,5

We now discuss the situation of author B. We use $\#(\mathrm{N2_B})$ to denote the number of combinations of these numbers of citations of the original papers in the h-core and two new papers of author B. These two new papers also receive $\#(\mathrm{C_{21}})$ and $\#(\mathrm{C_{22}})$ citations.

$$\#(\mathrm{N2_B}) = \binom{\mathrm{M_F} - \mathrm{h_B} + 1 + h_\mathrm{B} + 2 - 1}{h_\mathrm{B} + 2} - \binom{\mathrm{h_A} - \mathrm{h_B} + h_\mathrm{B} + 2 - 1}{h_\mathrm{B} + 2}$$
$$= \binom{\mathrm{M_F} + 2}{h_\mathrm{B} + 2} - \binom{\mathrm{h_A} + 1}{h_\mathrm{B} + 2} = \binom{9}{5} - \binom{6}{5} = 126 - 6 = 120$$

These 120 arrays are shown in table 6

Table 6. Arrays of the numbers of citations of the papers in the h-core and two new papers of author B

7,7,7,7,7	7,7,6,5,4	7,6,6,5,4	7,5,5,3,3	6,6,5,5,5	6,4,4,4,4
7,7,7,7,6	7,7,6,5,3	7,6,6,5,3	7,5,4,4,4	6,6,5,5,4	6,4,4,4,3
7,7,7,7,5	7,7,6,4,4	7,6,6,4,4	7,5,4,4,3	6,6,5,5,3	6,4,4,3,3
7,7,7,7,4	7,7,6,4,3	7,6,6,4,3	7,5,4,3,3	6,6,5,4,4	6,4,3,3,3
7,7,7,7,3	7,7,6,3,3	7,6,6,3,3	7,5,3,3,3	6,6,5,4,3	6,3,3,3,3
7,7,7,6,6	7,7,5,5,5	7,6,5,5,5	7,4,4,4,4	6,6,5,3,3	5,5,5,5,5
7,7,7,6,5	7,7,5,5,4	7,6,5,5,4	7,4,4,4,3	6,6,4,4,4	5,5,5,5,4
7,7,7,6,4	7,7,5,5,3	7,6,5,5,3	7,4,4,3,3	6,6,4,4,3	5,5,5,5,3
7,7,7,6,3	7,7,5,4,4	7,6,5,4,4	7,4,3,3,3	6,6,4,3,3	5,5,5,4,4
7,7,7,5,5	7,7,5,4,3	7,6,5,4,3	7,3,3,3,3	6,6,3,3,3	5,5,5,4,3
7,7,7,5,4	7,7,5,3,3	7,6,5,3,3	6,6,6,6,6	6,5,5,5,5	5,5,5,3,3
7,7,7,5,3	7,7,4,4,4	7,6,4,4,4	6,6,6,6,5	6,5,5,5,4	5,5,4,4,4
7,7,7,4,4	7,7,4,4,3	7,6,4,4,3	6,6,6,6,4	6,5,5,5,3	5,5,4,4,3
7,7,7,4,3	7,7,4,3,3	7,6,4,3,3	6,6,6,6,3	6,5,5,4,4	5,5,4,3,3
7,7,7,3,3	7,7,3,3,3	7,6,3,3,3	6,6,6,5,5	6,5,5,4,3	5,5,3,3,3
7,7,6,6,6	7,6,6,6,6	7,5,5,5,5	6,6,6,5,4	6,5,5,3,3	5,4,4,4,4
7,7,6,6,5	7,6,6,6,5	7,5,5,5,4	6,6,6,5,3	6,5,4,4,4	5,4,4,4,3
7,7,6,6,4	7,6,6,6,4	7,5,5,5,3	6,6,6,4,4	6,5,4,4,3	5,4,4,3,3
7,7,6,6,3	7,6,6,6,3	7,5,5,4,4	6,6,6,4,3	6,5,4,3,3	5,4,3,3,3
7,7,6,5,5	7,6,6,5,5	7,5,5,4,3	6,6,6,3,3	6,5,3,3,3	5,3,3,3,3

Change in h-index values of authors A and B

If each of author A and author B publishesthe same number of new papers, such that they receive the same numbers of citations (pairwise) and such that the numbers of citations are all larger than h_A, then what do the h-indices of both authors become?

We assume that the h-index of author A becomesh_{A1} and the h-index of author B becomesh_{B1} after each of the above authors published one new paper with the same number of citations. The h-index of author A becomesh_{A2}, the h-index of author B becomesh_{B2} after each of the authors published two more papers. The numbers of citations of these papers are all not smaller than h_A.

It is possible that the h-index will increase by one, and it is also possible that the h-index will not change after the author publishes one more paper receiving a number of citations larger than h. For the two authors with different h-indices, it is possible that the larger h-index increases by one, and the smaller h-index stays the same. It is also possible that the larger h-index does not change but the smaller h-index increase by one. It is also possible that the h-indices of both authors increase by one or stay the same. These possible changes of the h-indices are shown in table 7.

Table 7. Possible changes of the h-indices of author A and author B after adding some more papers

original h-indices	Adding one paper	$h_{A1} - h_{B1}$	Adding two papers	$h_{A2} - h_{B2}$
h_A,h_B	$h_A + 1, h_B$	$h_A + 1 - h_B$	$h_A + 2, h_B$	$h_A - h_B + 2$
			$h_A + 2, h_B + 1$	$h_A - h_B + 1$
			$h_A + 1, h_B$	$h_A - h_B + 1$
			$h_A + 1, h_B + 1$	$h_A - h_B$
	$h_A + 1, h_B + 1$	$h_A - h_B$	$h_A + 2, h_B + 1$	$h_A - h_B + 1$
			$h_A + 2, h_B + 2$	$h_A - h_B$
			$h_A + 1, h_B + 1$	$h_A - h_B$
			$h_A + 1, h_B + 2$	$h_A - h_B - 1$
	h_A, h_B	$h_A - h_B$	$h_A + 1, h_B$	$h_A + 1 - h_B$
			$h_A + 1, h_B + 1$	$h_A - h_B$
			h_A, h_B	$h_A - h_B$
			$h_A, h_B + 1$	$h_A - h_B - 1$
	$h_A, h_B + 1$	$h_A - h_B - 1$	$h_A + 1, h_B + 1$	$h_A - h_B$
			$h_A + 1, h_B + 2$	$h_A - h_B - 1$
			$h_A, h_B + 1$	$h_A - h_B - 1$
			$h_A, h_B + 2$	$h_A - h_B - 2$

$h_A - h_B > 0$, only when $h_{A1} - h_{B1} \leq 0$, the problem of inconsistencycan occur. Obviously the first step to reach $h_{A1} - h_{B1} \leq 0$ is that $h_{A1} - h_{B1} < h_A - h_B$.

If $h_A - h_B = 1$,both authors publish one more article, the number of citations of this paper is larger than h_A, then only when $h_{A1} - h_{B1} = h_A - h_B - 1 = 0$, i.e. the last row and the third column in the table 7do we have an inconsistency.
If $h_A - h_B = 2$, when both authors only publish one more paper, there is no problem of inconsistency. The problem of inconsistency only occurs when both authors publish two more papers:$h_{A2} - h_{B2} = h_A - h_B - 2 = 0$. This situation is shown in the last row and the fifth column in table 8.
When ish$_{A1} - h_{B1} < 0$? If $h_A > h_B$ and $h_A - h_B = 1$, there must at least be two new papers. When ish$_{A2} - h_{B2} < 0$? If $h_A > h_B$ and $h_A - h_B = 2$ there must at least be three new papers. Other situations are feasible; however, we do not go further. We just discuss how the smaller h-index catches up with the larger one.

Probability of inconsistency

We now use our example to calculate the probability thath$_{A1} - h_{B1} < h_A - h_B$ and the probability thath$_{A2} - h_{B2} = 0$when $h_A - h_B = 2$.
If $h_{A1} - h_{B1} < h_A - h_B$, then the h-index of author A must stay the same and the h-index of author B must increase by one when the two authors publish one more paper.
If $h_{A2} - h_{B2} = 0$, then the h-index of author A must stay the same and the h-index of author B must increase by two after publishing each two more papers.
We first calculate the probability of $h_{A1} - h_{B1} < h_A - h_B$ when both authors publish one new paper,and then the probability of $h_{A2} - h_{B2} = 0$ when both authors publish two new papers.

Author A as well as author B publish one newpaper

There are $h_A + 1$ integers in the arrays formed by the numbers of citations of the original papers in the h-core and one new paper of author A. If the h-index of author A didn't change after the author A published one more article, then the integer in the $(h_A + 1)^{th}$ position must be smaller than $h_A + 1$, since all numbers of citations of these papers are not smaller than h_A, so h_A must be placed in this position. The other h_A integers are selected from the set of integers in the interval $[h_A, M_F]$.
How many arrays can make $h_{A1} = h_A$? This is equivalent to acombinatorial problem in how many different ways h_A integers can be repeatedly selected from the set of integers in the interval $[h_A, M_F]$.
We use $P1_{(h_{A1}=h_A)}$ to denote the probability of $h_{A1} = h_A$and use $\#(N1_{(h_{A1}=h_A)})$ to denote the number of arrays that make $h_{A1} = h_A$.

$$\#(N1_{(h_{A1}=h_A)}) = \binom{M_F - h_A + 1 + h_A - 1}{h_A} = \binom{M_F}{h_A} = \binom{7}{5} = 21$$

$$P_{(h_{A1}=h_A)} = \frac{\#\left(N_{I(h_{A1}=h_A)}\right)}{\#(N_{IA})} = \frac{\binom{M_F}{h_A}}{\binom{M_F+1}{h_A+1}} = \frac{(h_A+1)}{(M_F+1)} = \frac{21}{28} = 0.75$$

We then discuss the probability of $h_{B1} > h_B$ after author B published one more paper. The number of citations this paper has received is larger than h_A.

We use $P1_{(h_{B1}>h_B)}$ to denote the probability of $h_{B1} > h_B$. We use $\#\left(N1_{(h_{B1}>h_B)}\right)$ to denote the number of arrays that make $h_{B1} > h_B$.

$$\#\left(N1_{I(h_{B1}>h_B)}\right)$$
$$= \binom{M_F - (h_B+1) + 1 + h_B + 1 - 1}{h_B+1} - \binom{h_A - (h_B+1) + h_B + 1 - 1}{h_B+1}$$
$$= \binom{M_F}{h_B+1} - \binom{h_A - 1}{h_B+1} = 35 - 1 = 34$$

$$P_{I(h_{B1}>h_B)} = \frac{\#\left(N_{I(h_{B1}>h_B)}\right)}{\#(N_{IB})} = \frac{\binom{M_F}{h_B+1} - \binom{h_A-1}{h_B+1}}{\binom{M_F+1}{h_B+1} - \binom{h_A}{h_B+1}} = \frac{34}{65} = 0.523077$$

We now discuss the probability of $h_{A1} = h_A$ and $h_{B1} > h_B$. We use $P1_{(h_{A1}=h_A, h_{B1}>h_B)}$ to denote the probability of $h_{A1} = h_A$ and $h_{B1} > h_B$.

$$P1_{(h_{A1}=h_A, h_{B1}>h_B)} = \frac{\binom{M_F}{h_A}}{\binom{M_F+1}{h_A+1}} * \frac{\binom{M_F}{h_B+1} - \binom{h_A-1}{h_B+1}}{\binom{M_F+1}{h_B+1} - \binom{h_A}{h_B+1}} = \frac{(h_A+1)}{(M_F+1)} * \frac{\binom{M_F}{h_B+1} - \binom{h_A-1}{h_B+1}}{\binom{M_F+1}{h_B+1} - \binom{h_A}{h_B+1}}$$
$$= \frac{21}{28} * \frac{34}{65} = 0.392308$$

This is a case of potential probability of inconsistency. The larger h-index is still larger than the smaller, though the advantage of author A has become smaller.

Both author A and author B publish two new papers

There are $h_A + 2$ integers in the arrays formed by the numbers of citations of the original papers in the h-core and two new papers of author A.

If the h-index of author A didn't change after author A published two more articles, then the integer in the $(h_A+1)^{th}$ position must be smaller than $h_A + 1$, since all numbers of citations of these papers are not smaller than h_A, so h_A must be placed in this position. The first h_A integers are from the set of integers in the interval $[h_A, M_F]$. How many arrays can make $h_{A2} = h_A$?This is equivalent to the combinatorial problem of finding in how many different ways h_A integers can be repeatedly selected from the set of integers in the interval $[h_A, M_F]$.

We use $P2_{(h_{A2}=h_A)}$ to denote the probability of of $h_{A2}=h_A$and use $\#(N2_{(h_{A2}=h_A)})$ to denote the number of arrays that make $h_{A2}=h_A$.

$$\#(N2_{(h_{A2}=h_A)})=\binom{M_F-h_A+1+h_A-1}{h_A}=\binom{M_F}{h_A}=\binom{7}{5}=21$$

$$P2_{(h_{A2}=h_A)}=\frac{\#(N2_{I(h_{A2}=h_A)})}{\#(N2_{IA})}=\frac{\binom{M_F}{h_A}}{\binom{M_F+2}{h_A+2}}=\frac{21}{36}=0.58333$$

We use $P2_{(h_{B2}>h_{B1})}$ to denote the probability of $h_{B2}>h_{B1}$ and use $\#(N2_{I(h_{B2}>h_{B1})})$ to denote the numbers of arrays that make $h_{B2}>h_{B1}$.

$$\begin{aligned}&\#(N2_{I(h_{B2}>h_{B1})})\\&=\binom{M_F-(h_B+2)+1+h_B+2-1}{h_B+2}-\binom{h_A-(h_B+2)+h_B+2-1}{h_B+2}\\&=\binom{M_F}{h_B+2}-\binom{h_A-1}{h_B+2}=\binom{7}{5}-\binom{4}{5}=21\end{aligned}$$

$$P2_{I(h_{B2}>h_{B1})}=\frac{\#(N2_{I(h_{B2}>h_{B1})})}{\#(N2_{IB})}=\frac{\binom{M_F}{h_B+2}-\binom{h_A-1}{h_B+2}}{\binom{M_F+2}{h_B+2}-\binom{h_A+1}{h_B+2}}=\frac{21}{120}=0.175$$

We now discuss the probability of $h_{A2}=h_A$ and$h_{B2}>h_{B1}$.

We use $P2_{(h_{A2}=h_A,h_{B2}>h_{B1})}$ to denote the probability of $h_{A2}=h_A$ and $h_{B2}>h_{B1}$$P2_{(h_{A2}=h_A,h_{B2}>h_{B1})}=\frac{\binom{M_F}{h_A}}{\binom{M_F+2}{h_A+2}}*\frac{\binom{M_F}{h_B+2}-\binom{h_A-1}{h_B+2}}{\binom{M_F+2}{h_B+2}-\binom{h_A+1}{h_B+2}}=\frac{21}{36}*\frac{21}{120}=0.102083$

We list all our calculations in table 8

Table 8. Probability of inconsistency. (The maximum number of citations is 7, the h-index of author A is 5, and the h-index of author B is 3).

	Author	Total number of arrays	Number of arrays that might give rise to inconsistency	Proportion	Probability of inconsistency
publishes one paper	A	28	21	0.75	0.392308
	B	65	34	0.523	
publishes two papers	A	36	21	0.583	0.102083
	B	120	21	0.175	

Using the same method we also calculate the probability of inconsistency of the h-index when the difference of the h-indices is 1, still assuming that the maximum number of citations is 7. So the larger h-index changes from 2 to 7, and the smaller one changes from 1 to 6. The results are shown in table 9.

Table 9. Comparison of the probability of inconsistency between different h-indices

	Authors	Total number of arrays	Number of arrays that might give rise to inconsistency	Proportion	Probability of both authors
$h_A = 2$	A	56	21	0.375	0.291667
$h_B = 1$	B	27	21	0.778	
$h_A = 3$	A	70	35	0.5	0.318182
$h_B = 2$	B	55	35	0.636	
$h_A = 4$	A	56	35	0.625	0.317029
$h_B = 3$	B	69	35	0.507	
$h_A = 5$	A	28	21	0.75	0.286364
$h_B = 4$	B	55	21	0.381	
$h_A = 6$	A	8	7	0.875	0.334091
$h_B = 5$	B	55	21	0.382	
$h_A = 7$	A	1	1	1	0.142857
$h_B = 6$	B	7	1	0.143	

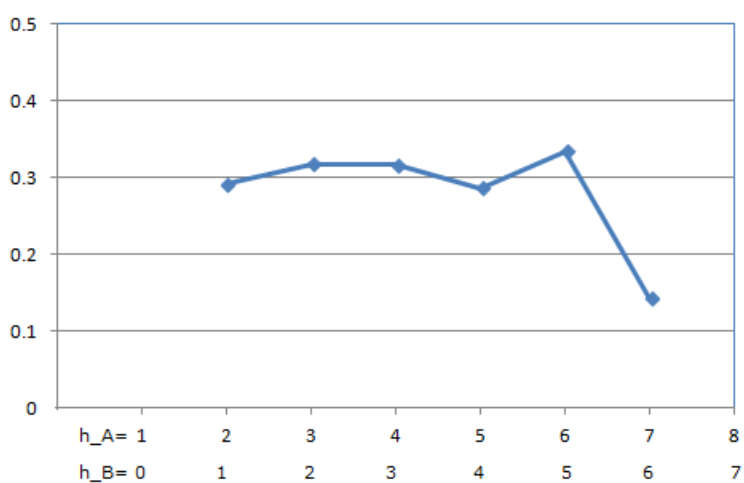


Figure 1. Change of the probability of inconsistency according to the h-index

From figure 1, we can see that the probability of inconsistency is increasing when the h-index is small, it reaches a maximum when the h-index is in the middle between 0 and the maximum number of citations, then decreases, but in the end, it fluctuates again. It reaches a minimum when the larger h-index is equal to the maximum number of citations.

Conclusion and Discussion

We illustrated the fact that the probability of inconsistency of the h-index can be calculated by combinatorial theory. We showed the factors such as the h-indices of two authors, the difference between the h-indices of the authors, the maximum numbers of citations of that two authors can receive, the number of the new papers the authors publish and the numbers of citations these new papers receive, that influence the probability of inconsistency of the h-index.

We still did not investigate how these factors influence the extent of the inconsistency in detail. Intuitively, the larger the maximum number of citations and the bigger the difference between the h-indices of the authors is, the smaller the probability of inconsistency. Our preliminary calculations confirm this point, i.e., the probability of inconsistency of the h-index is smaller when the difference of the h-indices of authors is 2 than when the difference of the h-indices is 1. However this aspect should be further investigated so that we know exactly, or at least approximately, how the probability of inconsistency changes according to the h-indices of the authors and the maximum number of citations. It should also be checked if the extent of inconsistency can be tolerated. If 0.1probability of inconsistency cannot be tolerated, then how about 0.0001? How can the h-index and the maximum number of citations of each author lead to such a low probability of inconsistency?

What does inconsistence mean to the development of science or to the progress of a scientist? We observe that only when the number of papers that two authors publish later is not smaller than the difference of the h-indices of these two authors, and the numbers of citations that these papers have received is larger than or equal to the h-index of author A, i.e., the higher h-index of two authors, the problem of inconsistency may occur. Can we think of inconsistency as a sign that author B is making progress? If the number of citations really is a scientific standard, then author B has published some more papers and the papers have received so many citations that these papers can enter the h-core of author A. Why is it not a sign that author B is catching up with author A? Consider two authors, the first one has 1000 papers with 1000 citations each. The second has only one paper with one citation, the h-index of the first author is 1000 whereas the h-index of the second is 1. When each of the above authors publishes 1000 more papers with 1000 citations, the h-index of the first author is now still 1000, and the h-index of the second author also reaches 1000, why can we not regard such an event as showing that the second author is making great progress and the first author just publishes on the same level?

Is consistency a necessary requirement when scientific indicators are designed to measure the development of science or the progress of a scientist? Our world is diverse, and hence how things in the world develop is also diverse. When the world is always consistent, the trendsetter is always a trendsetter, the laggard is always a laggard, and all scientists behave as the soldiers in parade formation. Is this possible? If the laggard makes some progress, we certainly should say he

became a better scientist. How does a laggard catch up with the trendsetter? Certainly if he does much better than the trendsetter, then we say he is making such a progress that he is now as good as or even better than the trendsetter. However, if he does as well as the trendsetter, it is also progress. Isn't it? Progress is made in different ways, science also develops in different ways, not in a consistent way. Scientific indicators designed to measure the development of science and the progress of a scientist should tolerate these differences.

Acknowledgement

This work was prepared when I was dealing with a crisis of life. I sincerely thank for the support from all members of the Class 1987 of Agricultural Information Science in Nanjing Agricultural University, Lu Junxi, Liang Jin, Du Juan, Wang Haiying, Yan Sulan, Zhang Min, Yin Jiahui, Wei Wenjie, Yao Junlan, QiuXuejun, Xu Wenyi, Dong Xiaoyun and many blog friends who encouraged me to pass through such a tough and tedious period. I thank Cao Cong for tolerating the delay of a collaborative work again and again. I also thank Raf Guns and Ronald Rousseau for valuable comments on an early version of this work. This work is supported by the National Natural Science Foundation of China (NSFC grant No. 71173154).

References

Bouyssou, D., & Marchant, T. (2010). Consistent bibliometric rankings of authors and journals. Journal of Informetrics, 4, 365–378.

Bouyssou, D., & Marchant, T. (2011a). Bibliometric rankings of journals based on impact factors: An axiomatic approach. Journal of Informetrics, 5(1),75–86.

Bouyssou, D., & Marchant, T. (2011b). Ranking scientists and departments in a consistent manner. Journal of the American Society for Information Science and Technology, 62(9), 1761–1769.

Hirsch, J. E. (2005). An index to quantify an individual's scientific research output. Proceedings of the National Academy of Sciences of the United States of America, 102(46), 16569–16572.

Liu, Y. X., & Rousseau, R. (2009). Properties of h-type indices: The case of library classification categories. Scientometrics, 79(2), 235–248.

Marchant, T. (2009a). An axiomatic characterization of the ranking based on the h-index and some other bibliometric rankings of authors. Scientometrics, 80(2), 325–342.

Marchant, T. (2009b). Score-based bibliometric rankings of authors. Journal of the American Society for Information Science and Technology, 60(6), 1132–1137.

Rousseau, R. (2012). Basic Properties of Both Percentile Rank Scores and the I3 Indicator. Journal of the American Society for Information Science and Technology, 63(2), 416-420.

Rousseau, R. (2011). Percentile rank scores are congruous indicators of relative performance, or aren't they? arxiv.org/pdf/1108.1860

Waltman, L. & van Eck, N.J. (2009a). A taxonomy of bibliometric performance indicators based on the property of consistency. In B. Larsen & J. Leta (Eds.), Proceedings of the 12th International Conference on Scientometrics and Informetrics (ISSI 2009) (pp. 1002–1003). Sao Paulo, Brazil: BIREME & Federal University of Rio de Janeiro.

Waltman, L., & van Eck, N.J. (2009b). A simple alternative to the h-index. ISSI Newsletter, 5(3), 46–48.

Waltman, L., & van Eck, N.J. (2011). The inconsistency of the h-index. Availableat: http://arxiv.org/ftp/arxiv/papers/1108/1108.3901.pdf [refertopublishedarticle]

Waltman, L., van Eck, N.J., van Leeuwen, T.N., Visser, M.S., & van Raan, A.F.J. (2011a). Towards a new crown indicator: Some theoretical considerations. Journal of Informetrics, 5(1), 37–47.

Waltman, L., van Eck, N.J., van Leeuwen, T.N., Visser, M.S., & van Raan, A.F.J. (2011b). Towards a new crown indicator: An empirical analysis. Scientometrics, 87(3), 467–481.

TOWARD A TIME-SENSITIVE MESOSCOPIC ANALYSIS OF CO-AUTHOR NETWORKS: A CASE STUDY OF TWO RESEARCH SPECIALTIES

Theresa Velden[1], Syed Ishtiaque Ahmed[2], Scott Allen Cambo[2], and Carl Lagoze[1]

[1] *tvelden, clagoze @umich.edu*
School of Information, University of Michigan, Ann Arbor, 105 S. State Street, Ann Arbor, MI 48109 (USA)

[2] *sa738, sac355 @cornell.edu*
Dept of Information Science, Cornell University, 301 College Avenue, Ithaca, NY 14850 (USA)

Abstract
In this paper we extend the mesoscopic analysis of scientific collaboration networks introduced in Velden et al. (2010) by adding a temporal dimension. We explore the temporal evolution of collaboration networks over a twenty-year period in two research specialties in the chemical and physical sciences. As a first step we compare global characteristics of the evolution of our networks with co-author networks covered in the recent literature to gauge their correspondence to those networks and assess the impact of data pre-processing steps that we perform on these global characteristics. Based on this comparison, we confirm (Milojevic 2010), but dispute the findings of (Abbasi et al 2012) of evidence suggesting a preferential attachment mechanism at work that would explain the evolution of network structure. We then turn to studying the growth and evolution of network structure at the level of connected components, and at the level of individual authors and groups of authors joining the network. Both fields we study experience a period of steady linear growth between 1996-2005, and the dominant social mechanism is the entry of independent new groups into the field that initially do not collaborate with existing groups in the field. Another important, similarly strong mechanism is that of junior researchers joining existing research groups in the field. Initial results indicate subtle field differences that require further study.

Conference Topic (see http://issi2013.org/about.html).
Collaboration Studies and Network Analysis (Topic 6) and Visualisation and Science Mapping: Tools, Methods and Applications (Topic 8)

Introduction
This paper reports on our latest results in a multiyear project that employs a mixed network analytic and ethnographic approach to understand the factors underlying field-specific attitudes towards openness and sharing of scholarly data. We report initial results of adding a temporal dimension to an analysis of scientific collaboration networks that provide evidence for comparative study of community structures and collaboration patterns across scientific fields. The

addition of a temporal dimension to the analysis allows us to study the dynamic processes involved in the evolution of a scientific community and to determine field specific patterns.

This work aims at advancing an ethnographically grounded approach to the mesoscopic analysis of collaboration networks (Velden et al. 2010, Velden & Lagoze 2013). By introducing a temporal dimension to the mesoscopic analysis of scientific networks we pursue the following two related goals. First we hope to increase the accuracy of community structure resolution to guide the strategic sampling of ethnographic field sites and interview partners in qualitative research (Velden & Lagoze 2013). Further, we aim to improve the validity of models that explain global evolution of network structures by linking dynamic features of network growth to specific social processes that can be verified in ethnographic field studies.

Previous work has oftentimes conceptualized co-author nodes as autonomous actors driven by individualistic mechanisms such as preferential attachment, ignoring the actual social composition of research collectives and the various socially distinct processes contributing to global network growth and densification. Supported by ethnographic insights, we can connect mesoscopic network features to notions of research groups, group leadership and implied seniority, inter-group collaboration, between-group migration, and ephemeral one-off exchanges. The promise of a mesoscopic approach is to support the interpretation of network dynamics in terms of distinct, superimposed social processes that can be verified and understood by ethnographic observation, and hence should allow us to significantly improve the validity of models to explain network evolution.

As first steps toward a time-sensitive mesoscopic analysis of co-author networks, we investigate in this study the following research questions:

How do global network metrics and specific results on network growth for our two datasets compare to the characteristics observed in other co-author networks recently studied? Our initial goal is to check that the data sets and the fields that we study are not outliers among other fields and data sets studied in the recent literature. We also want to assess the influence of pre-processing, data-cleaning steps such as author name disambiguation and hyper-authorship extraction, described in the Methods & Data section below, on global network metrics. For comparison we chose the following recent studies of co-author network growth and the evolution of its structure: two studies that investigate how authors join and establish new links in co-author networks (Abbasi et al. 2012, Milojevic 2010), and a comparative study of the evolution of the giant component in eight scientific fields (Bettencourt et al. 2009).

How do the co-author networks in the two research specialties evolve at the level of connected network components, and in particular in terms of the evolution of a giant component and network? We consider both accumulative network growth and dynamic network growth. In an accumulative scheme, co-author links and author nodes are added to the network in 1-year time steps over the entire 20-year range of our data, without ever removing any links or nodes. Accumulative growth has been the basis of most previous studies of the evolution of network topology (e.g. Newman 2001, Barabasi et al 2002). By also studying dynamic growth, we explore the evolution of network structure at a point in time, where 'point in time' refers to a time window of fixed size, aggregating co-author activity over a certain number of years and then moving that window along the time axis in 1-year steps. We document and compare between the two fields the flows between major components in the accumulative and dynamic scheme.

How do individual authors join the network? As we will discuss below, preferential attachment models that do not consider the social structure of research specialties and its organization in research groups fail to explain the evolution of co-author networks. To understand the specific social processes by which new authors enter a research specialty (e.g. a student joining a group already active in the field, or as a member of a research group that newly becomes active in this particular research specialty) we set out to identify network patterns that may represent specific entry scenarios and to quantify them. This allows us to assess their prevalence and to compare them across fields to look for commonalities or potentially significant field differences.

Methods & Data

We are developing an open source code base (http://github.com/tvelden/communities) that allows us to flexibly generate co-author networks following different time-slicing schemes: 'accumulative' for tracking the accumulative growth of the network, and 'sliding' for generating a dynamic view of the evolution of network structures by considering only publications in a specific time window. This sliding window can move across the entire time range covered by the available data. We have integrated methods into the code that support the mesoscopic analysis of networks, such as the network clustering code by Rosvall & Bergstrom (2008) and our own implementation of a node classification algorithm for clustered networks by Guimera et al (2007). This classification scheme allows us to distinguish types of nodes by their structural embedding into their surrounding co-author cluster and by their out of cluster connectivity. For example, hub nodes extracted from our networks by this classification scheme typically correspond in real life to research group leaders (Velden et al. 2010).

We employ lexical queries that extract from the Web of Science (WoS) of Thomson Reuters the publication output of two research specialties in the

physical and chemical sciences between 1991- 2010, one in synthetic chemistry (field 1), and one at the boundary of physics and physical chemistry (field 2).

Table 1. Basic Network Properties

	# papers	*# authors*	*# edges (weighted)*	*# edges (unweighted)*	*time period*
Field 1	12,641	13,397	58,375	31,858	1991-2010
Field 2	56,122	60,457	315,491	166,203	1991-2010

An important step in our analysis is the cleaning of data. An initial step is the *normalization* of author names. We capitalize names and concatenate hyphenated names. Next, to improve the accuracy of the co-author networks, we apply an *author name disambiguation* algorithm that has been shown to improve the resolution of individual authors. This step removes misleading distortions in the network structure due to name homonymy (Velden et al. 2011). We further use a statistical approach to define *hyper-authorship* in a data set-specific way[169] and use it to exclude a small set of papers (1-3%) that are not representative of the research style in the long-tail science fields that we study here. A manual analysis finds that in many cases those hyper-authorship papers represent out-of-scope papers that the lexical query mistakenly captured. In a few cases we also find large-scale collaborations that contribute to the specific field we study, but represent only a marginal sub-community within the field. Finally, we *exclude authors who have co-authored only a single paper*. About ⅔ of authors are removed in this step. The reduced data is much more manageable for analysis and visualization purposes. The effects of these reduction steps on network topology are reported in the results section. In the following we will use the labels 'norm' to refer to the data that has been merely name normalized, 'norm-dis' to data that has been normalized and disambiguated, 'norm-dis-hfree' for data that has been additionally filtered to exclude hyper-authorship papers, and finally 'norm-dis-hfree-red' for the data that has undergone all four preprocessing steps and represents our preferred data set for future analyses.

Results

Here we report the results of our network analysis of data in two research specialties, following the list of research questions outlined in the introduction.

Comparison of Global Network Metrics

We compare the following global network metrics with those obtained by Bettencourt et al. in their study of eight scientific fields (Bettencourt et al. 2009): the relative size of the giant component, the evolution of the diameter of the giant

[169] Given the long-tail distribution of co-authors over papers, we use a non-parametric approach to identify hyper-authorship papers as outliers based on median absolute deviation (described in http://rfd.uoregon.edu/files/rfd/StatisticalResources/outl.txt).

component defined as the maximum shortest distance between any pair of nodes in the network, and the scaling parameter of network densification.

As shown in table 2, we find that our network data represent scientific fields within a typical range of topological characteristics. In particular, for both of our fields, the scaling parameter for network densification is greater than 1.0 and similar e.g. to the field of carbon nanotubes. Following the argument by Bettencourt et al., this is an indication that we are dealing with 'non-pathological' fields, that is a community of researchers that share concepts and techniques, which facilitates collaboration. Further, it is likely that the network metrics of our data would be even more similar to Bettencourt et al. if their data were preprocessed in the same way[170,171]. This trend is indicated by the results we obtain for our only minimally treated data ('norm') in table 2.

Table 2. Comparison of Global Network Metrics

	giant component (% edges)	*diameter giant component*	*scaling parameter network densification*
Field1 (norm-dis-hfree-red)	~ 77	~ 34	1.14
Field1 (norm)	~ 87	~ 30	1.13
Field2 (norm-dis-hfree-red)	~ 86	~ 42	1.18
Field2 (norm)	~ 92	~ 29	1.27
String theory	~ 90	16	1.36
Carbon Nanotubes	~ 95	12-14	1.17

Influence of Data Cleaning on Global Metrics

The influence of the data cleaning steps on global network metrics is documented in table 3. Both fields show similar trends. The overall effects of cleaning the data on the global metrics of the resulting co-author networks are the following:

The *relative giant component size* of the final network is around 10% smaller than the giant component size derived from the initial data. The *diameter of the giant component* increases with the processing by about 25% (field 1) and 34% (field 2). The *scaling parameter of field densification* initially decreases and then increases again for the various pre-processing steps. All values obtained are within the desirable range of > 1 and the fluctuations are modest, well within the

[170] The data used by Bettencourt et al. is non-disambiguated. Also, whereas we consider two author name initials where provided, they consider only first initials of author names thereby worsening the problem of name homonymy. Further their data processing does not include a reduction step like the one we used.

[171] The difficulty comparing results across studies by different authors highlights the desirability of data sharing within the scientometric community, which would facilitate cross-validation and build a sound empirical basis for observations across research fields.

range of differences found between different fields within Bettencourt et al's analysis of eight scientific fields. The *peak of the degree distribution* (Milojevic 2010) remains at three for both fields until the final reduction step, when it falls to 2 after all 1-paper authors have been removed from the data set. Finally, *the proportion of authors that constitute the hook* part of the distribution (roughly all authors with less than 9 collaborators for our data) increases through the data cleaning process by about 10%. The biggest increase is caused by the removal of hyper-authorship papers, presumably since in this step a significant number of authors that contributed to the long tail of the distribution were removed.

Table 3. Influence of Pre-processing Steps on Global Metrics

	norm	*norm-dis*	*norm-dis-hfree*	*norm-dis-hfree-red*
Giant size (% edges)				
Field 1	86.9	69.2	67.5	76.7
Field 2	92.6	76.2	71.4	84.2
Giant Diameter				
Field 1	~29	~32	~33	~34
Field 2	~30	~36	~39	~41
Network Densification				
Field 1	1.13	1.12	1.10	1.14
Field 2	1.27	1.22	1.14	1.18
Peak of Degree Distrib.				
Field 1	3	3	3	2
Field 2	3	3	3	2
% Authors in 'Hook'				
Field 1	81.0	82.2	88.3	88.6
Field 2	72.6	76.7	86.1	84.0

As would be expected the data cleaning steps affect the numerical values of global network metrics. However, from these observations we conclude that the network based on cleaned data preserves the important topological features of the initial network. The numerical changes are below an order of magnitude, and characteristics such as the scaling parameter that Bettencourt et al. use to identify pathological fields remain within the acceptable range (> 1).

Correlation of node centrality measures with preferential attachment

We check whether we can reproduce for our data the results of Abbasi et al. (2012) regarding the role of preferential attachment as a mechanism driving network growth. Abbasi et al. compared three node centrality metrics (betweenness, degree, closeness) for their correlation with preferential attachment. They investigated the hypothesis that existing nodes in the network with higher centrality values attract a higher number of co-authors that are either 'newbies' (new in the field) or 'oldies' (existing members of the field that they

had not previously co-authored with). Their major finding was that betweenness centrality trumps degree centrality for attracting newbie authors, which traditionally had been the focus of preferential attachment models. However, in their data, degree centrality correlates stronger with the number of new collaborations with 'oldies' authors. Their data set covered a single field of research (steel structure research from 1999-2009) and they called for broader validation of this finding across further scientific fields.

In contrast to Abbasi et al. we cannot confirm for our data that betweenness centrality drives preferential attachment of newbie authors to existing authors in the field, outperforming degree centrality. Instead we obtain correlation values that are very similar for degree centrality and betweenness centrality (figure 1). Also our data show an even lower correlation strength for betweenness centrality than that reported by Abbasi et al. (between 0.15 and 0.25 for field 1, and 0.10 and 0.20 for field 2). We consider the 0.23 to 0.32 Spearman correlation strengths that Abbasi et al. report for the correlation of betweenness centrality with number of collaborations with newbie authors as relatively low, having limited explanatory value for the variation observed in the number of links existing authors form with new authors. Similar to Abbasi et al. we find that degree centrality has the highest correlation values for new co-author links between already existing authors. Those correlations are of medium strength (around 0.45 for both fields, figure 1).

We double-checked whether the deviation of our results from Abbasi et al's could be explained through the different pre-processing steps we applied to clean our data. We assume that the version of our data that includes name disambiguation and removes hyper-authorship papers but omits the last preprocessing step of removing 1-paper authors may be closest to the cleaning protocol that Abbasi et al. used. However, even for this data we cannot find confirmation of their result of the dominant role of betweenness centrality for preferential attachment of new authors to existing authors. Instead for our data in this treatment regime degree centrality slightly dominates over betweenness centrality.

Degree distribution and deviations from scaling law behavior (Milojevic 2010)

As pointed out by Milojevic (2010) power law scaling of the distribution of number of collaborators (associated with a hypothesized preferential attachment mechanism at work), is not the dominant feature characterizing such distributions in co-author networks. Instead, the majority of authors (88% in the 2000-2004 data set of nano-science publications studied by Milojevic) are included in the log-normal hook of the distribution. We find similar numbers for the proportion of authors constituting the hook, ranging between 70% and 90% depending on the field and data pre-processing steps used (table 3).

Milojevic interprets the hook and its peak as suggestive of a characteristic mode of collaboration corresponding to the typical number of collaborators needed in a research field to produce a publishable result. Our data for both fields display the same log-normal hook feature with a peak at 2 collaborators. These peak values are slightly smaller than those in nano-science subfields such as applied physics or materials chemistry that are intuitively comparable to our fields. As discussed above, this could be due to differences in pre-processing of the data. The data used by Milojevic is probably best compared to our initial data, i.e. the data set without name disambiguation, without removal of hyper-authorship, and without removing 1-paper authors. For this version of our data we obtain peaks at 3 collaborators for both fields (see table 3), the same Milojevic finds in 2000-2004 for the analytical chemistry subfield within nano sciences, but lower by one than found by her for the applied physics and the materials chemistry subfields within nano sciences.

Network growth at the component level

The focus in this result section is on how the network grows at the level of connected components, and in particular how the giant component of the network grows.

Initially we consider the accumulative view, i.e. the network as it grows by adding year-by-year new authors and new co-author links without removing any from previous years. When plotting the temporal evolution of the absolute size of the second largest component of the co-author network we noticed that the second largest component showed distinct reductions in size every few years. Within an accumulative scheme this phenomenon can only be explained by the second largest component having merged with the largest component, thereby allowing a previously smaller component to move up the ranks and become the second largest component even if smaller in size than the second largest component in the previous year. We investigated whether these 'feeding events' of the giant component by merging with the second largest component account for a substantial part in the growth of the giant component. We call this the *Staging Area Hypothesis*, i.e. the notion that the second largest component constitutes a major staging area for the giant component and that the largest component of a network gradually grows and evolves into the giant component of the network through subsequent absorption of the second largest component. Our subsequent analysis led to a rejection of this hypothesis for our data. We checked the contribution of the second largest component to the growth of the largest component and found that in none of our networks was the giant component primarily fed by the second largest component of the previous years (figure 2). New authors joining the field as well as authors that were previously part of components smaller than the second largest component provide for significantly stronger influx channels.

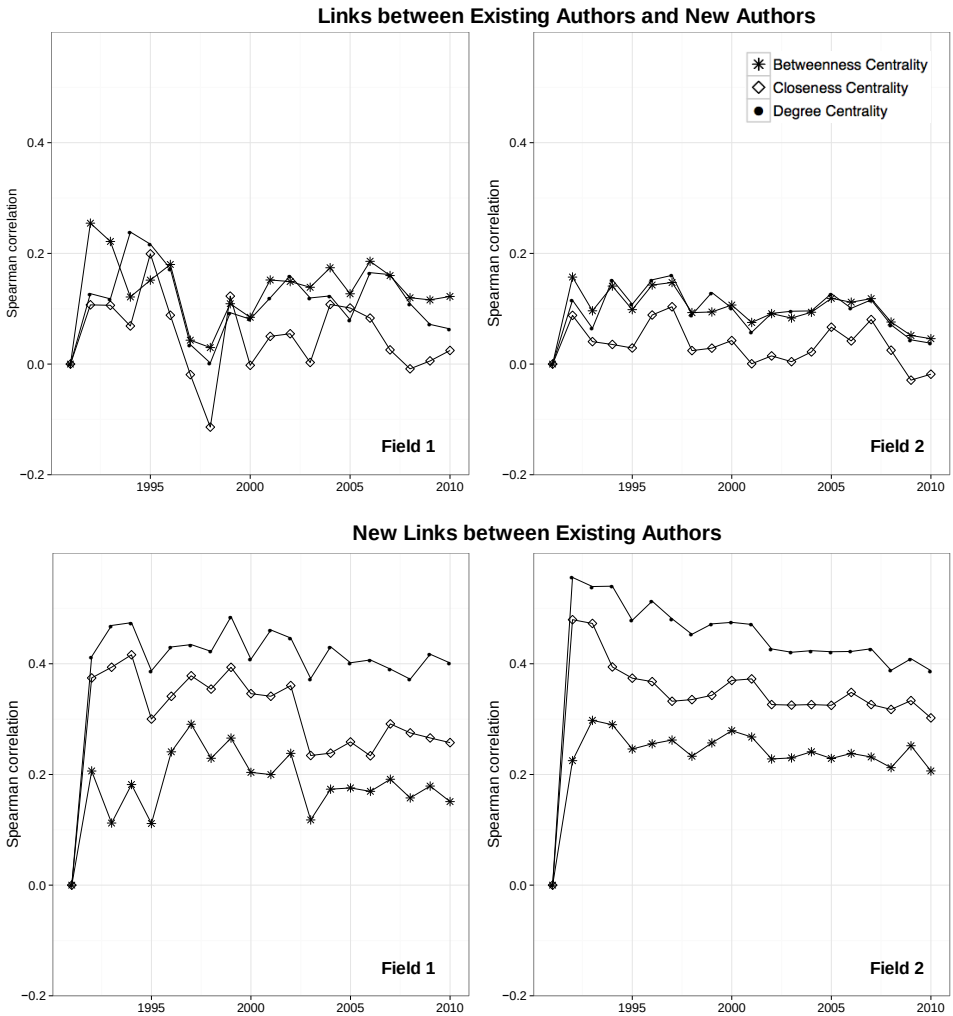


Figure 1: Correlation of node centrality scores of existing authors with co-author links to new authors (top) and with new co-author link to existing authors (bottom).

We also explore a dynamic perspective on the evolution of network components. We start by investigating the effect of window size on the size and evolution of the largest component of the co-author network. Publications are a fuzzy indicator of collaborative relationships, since they typically constitute only the end point of a successful research project, they lag somewhere between a few months up to several years behind the most intensive collaborative activity, and do not come out in strict chronological order of the underlying research activities. This suggests that one needs to accumulate data from a certain minimum number of years to capture the concurrent collaborative activities within a research group and its surrounding research community. We find for our data that for small sizes of the time window the largest component size fluctuates and mostly stagnates

over time. Only for a window size of 5 years and larger do we find stable growth of the giant component (for field 2 we observe a slight decline at the end of this period however). Hence, we chose a 5-year sliding window for our subsequent dynamic analysis.

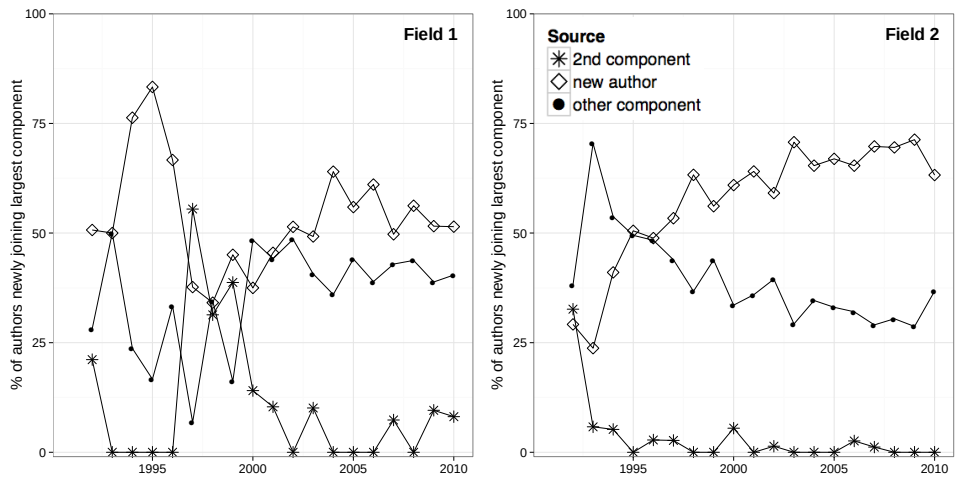


Figure 2: Source of nodes that newly join the giant component to check staging area hypothesis.

To visually explore the dynamic evolution of network components we use an online alluvial generator provided by Edler & Rosvall (2010) to show the temporal flow of authors between major components. This free online visualization tool generates the alluvial diagram from a chronologically organized series of networks. In their implementation, they basically show the flows among node clusters in a clustered network. However, in our case we were interested in the evolution of the connected components in the network. So, we generated a shadow-version of our networks that represent the components formally as clusters; and then we used the alluvial visualization tool to visualize the flows between the network components. We also considered the 'not yet active authors' in each time slice and represented them as an additional network component; these are authors who did not publish during the current network slice, or any of the previous slices, but who eventually published sometime in the future. The resulting visualization for field 1 is shown in figure 3, however the corresponding visualization for field 2 could not be obtained using the online tool available because of the size of the network.

As illustrated, the largest proportion of new ('not yet active') authors that get activated in a time step feed into the giant component. This contribution of new authors to the giant component dwarfs the contribution from continuing authors that were part of the giant component during the previous 5-year time window. This phenomenon is most pronounced for the growth period 1996-2000, and the

trend is similar, if less pronounced, for field 2. Consequently, new authors dominate the influx into the giant component even more so in the dynamic picture than the accumulative scheme.

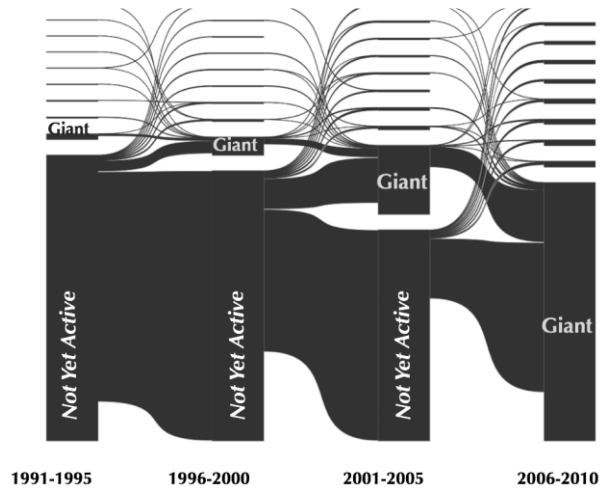


Figure 3: Alluvial flow of network nodes (authors) between network components for field 1 from a dynamic perspective

How New Authors Join A Research Specialty

The observations by Milojevic, as well as our own results that show the weak correlations between node centrality values and the attraction of new authors by existing authors, point to the weakness of a simple preferential attachment growth model to explain the evolution of co-author networks. Therefore we set out to have a closer look at how new authors join the network. To study the process by which the majority of new authors join the network each year, we go beyond a macroscopic approach that considers generic attachment patterns and break down the analysis to a mesoscopic level of analysis supported by ethnographic observations. Both field 1 and field 2 represent experimentally oriented research specialties in physics and chemistry. Their research communities are dominated by mid-sized research groups, typically composed of graduate students, postdoctoral associates and technicians, and led by a senior researcher (professor). Based on insights from ethnographic studies of these fields (Velden 2013) we can posit several realistic social processes through which new authors become active in these fields:

Scenario 1: a graduate student or a postdoctoral researcher joins a research group already publishing in the respective field and is trained within that group.
Scenario 2: as members of a research group that has been previously active in other areas and now starts to contribute to the research in the field independently as an autonomous actor

Scenario 3: a research group that enters the field by teaming up with a group already active in the field and contributes to the research in the field in collaboration with that group.

To derive a quantification of the relative importance of these social processes from the co-author networks and to compare them across fields, we develop an algorithmic procedure that attempts to map new authors entering the field each year to one of these scenarios based on their structural position within the 1-year co-author network in the year when they publish for the first time in the field, in combination with additional information derived from the 20-year accumulative network. A mesoscopic perspective provides us with the following network constructs that we make use of to operationalize and distinguish those processes:

Connected components: we distinguish connected components in the 1-year network based on whether they are composed solely of new authors, or solely of recurring authors, or of both.
Co-author Clusters: we extract the modular substructure of the accumulative 20-year co-author network[172] using an information theoretic algorithm (Rosvall & Bergstrom 2008) that subdivides the nodes in the co-author network into clusters that can be interpreted as research groups or collectives of very closely collaborating research groups (Velden et al. 2010).
Hub nodes: based on the structure of the clustered, accumulative co-author network, we can distinguish between hub nodes and non-hub nodes (Guimera et al. 2007). Hub nodes are defined as those nodes within a cluster of nodes that have a disproportionally high number of within-cluster links. As reported elsewhere (Velden et al. 2010), hub nodes within co-author networks are characteristic of the bibliographic footprint of research group leaders over time.

With the help of these constructs we operationalize the cases described in a two step procedure that eventually maps new nodes in the annual slice of the co-author network to one of the three scenarios. Nodes that cannot be mapped are assigned to a fourth scenario; essentially new nodes with an unresolved entry scenario. In a first step we parse each connected network component to assess its potential for representing the bibliographic footprint of one of the scenarios. Depending on the outcome of this assessment, the new nodes included in the component then get mapped either directly, or sometimes only after further analysis, to one of the four scenarios.

Given a network component that only consists of new authors, we map all these new authors to scenario 1. If a network component includes both new and recurring authors, we check for the presence of hub nodes. If we find in a component a linked pairs of hub nodes where one hub node is a new node and one

[172] This accumulative co-author network is constructed from publication data of the entire 20-year range from 1991-2010.

hub node is a recurring node, we consider the network component a candidate for representing scenario 3, i.e. a new research group entering the field through collaboration with an existing research group. However, before mapping all new nodes in this component to scenario 3, we evaluate whether the new hub node along with new non-hub nodes connected to it plausibly represent a new research group[173]. To this end we evaluate the relative connection strength between the new non-hub nodes with the new hub node and the existing hub nodes (including future years) to decide whether they can be counted as the research group of a research group leader newly entering the field or rather peers of a student who only later becomes a research group leader. For space limitations we here cannot provide the full algorithm specification. Instead it will be made available in the documentation of the source code available at http://github.com/tvelden/communities).

The results of our analysis are depicted in figure 4. We restrict our interpretation on a central 10-year time window (1996-2005) that represents a period of linear growth in both fields and should only be marginally affected by boundary effects[174]. We find that most authors join the field as part of new groups of authors becoming active in the field. However, there is also a relevant number of new authors who enter as new, presumably junior authors (students, postdocs) joining an existing group. Between 20-30% of new authors cannot be mapped to any of the cases (listed as scenario 4) since the interlinking patterns are too complex to be interpreted in an obvious way. With our approach we could determine only a very small number of new authors that enter the field as a group that immediately engages in collaboration with an existing group (and hence may be drawn into the field through collaboration with an existing group). We further observe that the two fields differ with regard to the percentage of nodes that can be mapped to either scenario 1 or cannot be resolved (scenario 4). Field 1 shows a higher percentage of scenario 1 authors whereas field 2 has a higher percentage of unresolved cases, consistent with our previous observation of the predominance of single hub clusters in field 1 and higher rates of inter-group collaboration and multiple hub clusters in field 2 (Velden et al. 2010). This may point to field

[173] The identification of hub nodes is based on clustering the 20-year accumulative co-author network. In a separate analysis we determined that it takes hub nodes about 5 years of publishing activity to expose the bibliographic footprint of a hub node. Hence a new author in the field that is categorized by us as a hub node may either be a research group leader from the beginning or a junior researcher (student, postdoc) who trains with an existing research group leader and later in her career becomes herself a research group leader active in this field.

[174] For this analysis we have to consider boundary effects at both ends of the time range covered by our entire data set. We can expect to overestimate the number of new authors and underestimate the number of recurring authors at the beginning of the time interval due to lack of knowledge about publishing activity of authors in the field before 1991. In an explorative analysis we found that this error is likely to reduce to less than 10% about 5-8 years into the time interval. At the end of the available time window we have issues with correctly recognizing hub nodes since the bibliographic footprint of a research group leader that enters the field 5 years or less before 2010 will remain too weak to be recognized as a hub node.

differences in the predominance of entry routes. However, it could also mean that the entry routes into the community are very similar in both fields but that the resolution of scenarios is made more difficult for field 2 because of the higher rate of inter-group collaborations. This question needs further exploration.

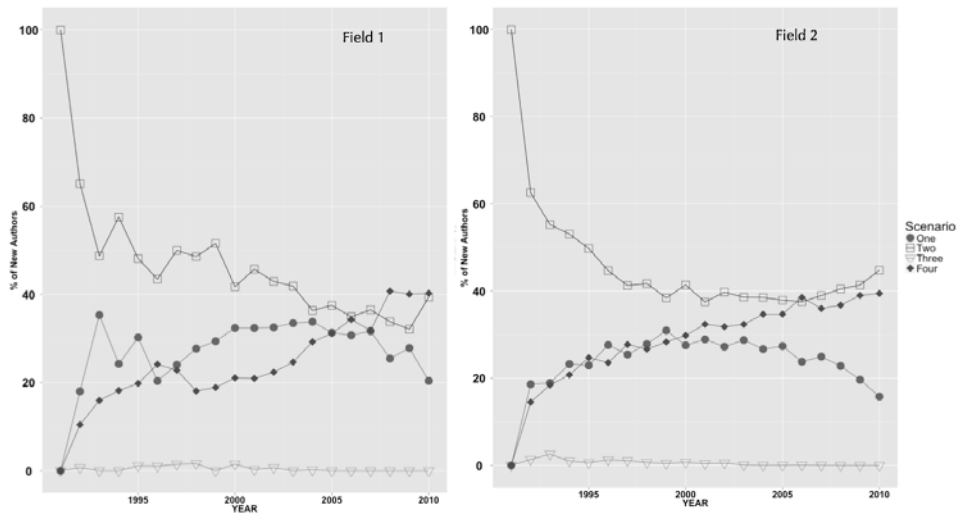


Figure 4: Quantification of social processes how new authors enter network each year. Scenario 1: authors enter the field by joining and existing research group. Scenario 2: authors enter as part of a new, independent research group. Scenario 3: authors enter field as members of a new research group that collaborates with an existing research group. Scenario 4: authors that we could not map to any of these cases.

Conclusions

We suggest that attempts at explaining the dynamics of co-author network growth in terms of realistic social scenarios need to distinguish carefully between the different types of nodes and processes underlying co-authorship collaborations. As seen, for example, from the analysis of the role of node centrality values in driving network growth, a generic preferential attachment mechanism has limited value for explaining the structural evolution of co-author networks. Instead, we believe that the evolution of collaboration networks in scientific communities and the nature of field-specific collaboration patterns would be more valuable if grounded by ethnographic observations. In this paper we report first steps for introducing time dimension into the analysis of co-author networks at the mesoscopic level. We focus on co-author networks, however in future work we anticipate including the temporal analysis of layered citation and co-author networks with the goal mapping of community structures within scientific fields (Velden & Lagoze 2013).

We note that our experiences with the replication of other authors' results revealed a number of critical issues that underline the potential benefit of an open data approach, allowing routine sharing of the data sets underlying published analyses, for developing a strong reliable empirical base for field comparisons.

Acknowledgments

National Science Foundation Grant No OCI-1025679, International Fulbright Science and Technology Fellowship for S.I. Ahmed.

References

Abbasi, A., Hossain, L., & Leydesdorff, L. (2012). Betweenness centrality as a driver of preferential attachment in the evolution of research collaboration networks. *Journal of Informetrics*, *6*(3), 403-412.

Barabâsi, A. L., Jeong, H., Néda, Z., Ravasz, E., Schubert, A., & Vicsek, T. (2002). Evolution of the social network of scientific collaborations. *Physica A: Statistical Mechanics and its Applications*, *311*(3), 590-614.

Bettencourt, L. M. a., Kaiser, D. I., & Kaur, J. (2009). Scientific discovery and topological transitions in collaboration networks. *Journal of Informetrics*, *3*(3), 210-221.

Edler, D. and Rosvall, M. (2010), The Map Generator software package, online at http://www.mapequation.org.

Guimera, R.; Sales-Pardo, M. & Amaral, L. (2007), 'Classes of complex networks defined by role-to-role connectivity profiles', *Nature Physics* **3**(1), 63-69.

Milojević, S. (2010). Modes of collaboration in modern science: Beyond power laws and preferential attachment. *Journal of the American Society for Information Science and Technology*, *61*(7), 1410-1423.

Newman, M. E. (2001). Clustering and preferential attachment in growing networks. *Physical Review E*, *64*(2), 025102.

Rosvall, M. & Bergstrom, C. (2008), 'Maps of information flow reveal community structure in complex networks', *PNAS* 105, 1118.

Velden, T.; Haque, A. & Lagoze, C. (2011), Resolving Author Name Homonymy to Improve Resolution of Structures in Co-author Networks, *in* 'JCDL'11, June 13-17, 2011, Ottawa, Ontario, Canada'.

Velden, T.; Haque, A. & Lagoze, C. (2010), 'A new approach to analyzing patterns of collaboration in co-authorship networks: mesoscopic analysis and interpretation', *Scientometrics* **85**(1), 2 19-242.

Velden, T. (2013), 'Explaining Field Differences in Openness and Sharing in Scientific Communities', CSCW 2013: Computer Supported Cooperative Work, February 23 - 27 2013, San Antonio, TX.

Velden, T. & Lagoze, C. (accepted), 'The Extraction of Community Structures from Publication Networks to Support Ethnographic Observations of Field Differences in Scientific Communication', JASIST

TOWARDS THE DEVELOPMENT OF AN INDICATOR OF CONFORMITY

Richard Klavans[1], Kevin W. Boyack[2], Aaron A. Sorensen[3], and Chaomei Chen[4]

[1] *rklavans@mapofscience.com*
SciTech Strategies, Inc., Berwyn, PA 19312 (USA)

[2] *kboyack@mapofscience.com*
SciTech Strategies, Inc., Albuquerque, NM 87122 (USA)

[3] *A.Sorensen@elsevier.com*
Elsevier, B.V., New York, NY 10017 (USA)

[4] *cc345@drexel.edu*
College of Information Science & Technology, Drexel University, Philadelphia, PA 19104 (USA)

Abstract
An indicator of conformity – the tendency for a scientific paper to reinforce existing belief systems – is introduced. This indicator is based on a computational theory of innovation, where an author's belief systems are compared to socio-cognitive norms. Evidence of the validity of the indicator is provided using a sample of 4180 high impact papers in two experiments. The first experiment is based on a 10 year model of the scientific literature. The robustness of the first experiment is tested using an alternative method for calculating the indicator and two 16-year models of the scientific literature.

Conference Topic
Scientometric Indicators (Topic 1)**;** Science Policy and Research Evaluation (Topic 3)

Introduction
For the past 50 years, there has been a working assumption in scientometrics that, overall, highly cited papers are innovative. As a consequence, there is a corresponding use of citation counts and other impact-related metrics to make resource allocation decisions in nations throughout the world (Geuna & Martin, 2003; Hicks, 2012; Martin, 2011). A university or institution is considered innovative if they have scientists that produce highly cited papers. National funding agencies are publicly criticized as being non-innovative if they didn't fund authors of highly cited papers (Nicholson & Ioannidis, 2012).

This assumption leads to significant questions that have not been considered. Specifically, what if many of these high impact papers are not really innovative, but simply reinforce the status quo? What might this policy of rewarding high impact do to the ability of a nation to fund highly innovative research? We simply

haven't looked at the possibility that high impact papers are reinforcing existing belief systems (i.e., our definition of conformity). Rewarding conformity will reduce the innovative output of a university or a nation.

The idea of an indicator of non-innovativeness, or conformity, is bound to be controversial because no one wants their work to be publicly recognized as not being innovative. On the other hand, there is a need to ensure that resources that are [hopefully] earmarked for innovative activities are not redirected towards institutions with a track record of non-innovative, but highly influential, research. This is a particularly difficult problem during a financial crisis where there is an across-the-board requirement to cut funding. It is politically easy to say no to a highly innovative (and potentially risky) program that hasn't been funded or where the researchers have little influence. It is much harder to cut the non-innovative program where the researchers are well established in a prestigious laboratory and have very high influence.

To the best of our knowledge, there is no scientometric research on high impact documents that are conforming. This study is a first step toward proposing an indicator of conformity. In the remainder of the paper we set the stage by reviewing related research. A preliminary experiment that was intended to develop an indicator of innovativeness (but unintentionally yielded potential information on conformity) is then described. A more comprehensive experiment intended to further explore the idea of conformity is elucidated. Finally, we summarize our results and point to further research that should be conducted.

Background

The idea that high impact papers are innovative is a commonly held belief. This was not always the case. This idea was extremely controversial in the 1960's when it was championed by Eugene Garfield. Garfield's idea was to create databases that listed scientific articles and their citations. These data became the basis of a commercial business (the Institute for Scientific Information or ISI) and became the organizing basis for the newly emerging academic field of scientometrics. Although retrieval was the primary purpose for introduction of the citation index, citation analysis soon became common, and was found to be a valid way to identify highly influential scientists, journals and papers. As evidence it was shown that Nobel Prize winners are among the most highly cited scientists (Garfield, 1977; Garfield & Malin, 1968; Garfield & Welljams-Dorof, 1992). Citation data were used to create a Journal Impact Factor (Garfield, 1972, 2006) that is now widely used. Between 1977 and 1993, Current Contents published over 4000 interviews with authors of citation classics describing the contributions of their highly cited papers. The cumulative evidence that highly cited papers are innovative seems overwhelming.

The possibility that high impact papers might not be innovative was brought up when citation analysis started. Most notable was the concern that a citation might be negative, perfunctory, or inflated due to self-citations (MacRoberts & MacRoberts, 1989, 1996; Moravcsik & Murugesan, 1975). These criticisms have been dealt with over time. There was no strong evidence that negative or perfunctory citations were systematically distorting impact indicators. A self-citation doesn't mean that a paper isn't innovative. Rather, it could simply signal that a researcher is building on his or her prior work. The only criticism that was generally considered legitimate was that review papers should be considered separately because of their unique role in science. Review papers were correspondingly identified and treated separately. They are typically excluded from ISI's list of high impact papers. What is left standing is the belief that highly cited papers, except for the review papers and a few exceptions, had to be innovative. This is such a widely held belief that a failure to fund high impact research was considered as legitimate sign of conformity by the editorial board of *Nature* (Nicholson & Ioannidis, 2012).

The possibility that high impact papers might be conforming is not an area of research in citation analysis. However, there is a significant amount of research on identifying innovative documents. It is from this stream of research that the proposed indicator was accidently developed.

A first experiment

The first experiment we designed built on the work of Chen et al. (2009) who used network analysis to identify innovative scientific papers. Chen posited that innovative papers would be located in the structural holes in citation networks. The scientific literature is envisioned as a network. Using co-citation analysis this network is composed of references, which are thought of as concept symbols that a newer paper builds upon. Typically, there are dense clusters of references that tend to be cited together. These dense clusters are separated from other dense clusters of references, and when visualized, look like islands in an archipelago. The essence of Chen's argument is that potentially innovative papers would tend to have high network betweenness. Using the analogy of islands, an innovative paper is more likely to appear somewhere between existing islands than in the center of an island. A new paper that cites multiple islands, or builds on the wisdoms of multiple islands, is more likely to be innovative than a paper predominantly referencing a single island.

Our intent was to conduct a large scale test of Chen's computational theory. Rather than duplicating their method, which operates on local datasets (one-at-a-time) and uses a relatively complex method for calculating network betweenness, we designed an experiment that would preserve their intent while using global models (in which the islands are all pre calculated) and simpler calculations.

We operationalized the notions of innovation (which correlates with betweenness) and conformity (which correlates with status quo, or lack of betweenness) using our global model as follows. Given a paper and its references, a pair of references that come from different clusters is a vote for betweenness or innovation. Conversely, a pair of references that come from the same cluster is a vote for status quo or conformity. We also needed to deal with missing information because our models do not necessarily include all references. In cases where one or both of the references were missing, we considered this as an undecided vote. In this experiment, for a given paper we counted votes from all possible pairs of references into three bins – innovative, conforming, or undecided.

In our study we also had the advantage of having a significant amount of full text information (2007 full text from Elsevier). It has been shown that proximate pairs of references (which can only be determined from full text) are more similar than pairs of references that are far apart in the text (Boyack, Small, & Klavans, 2013). With this information, one can test the effect of reference proximity on the votes – one can test whether more proximate pairs of references are better predictors of conformity or innovation than all pairs of references.

The most difficult issue in our experiment was in the design of the dependent variable. How do we know if we have identified an innovative paper? Chen's approach requires a significant amount of effort – measured in hours at least and perhaps days per turning point paper. We addressed this problem by looking at the stability of the flow of science using a global model of science. Our working hypothesis was that turning point papers would be in (a) unstable flows or (b) flows that are in increasingly unstable environments.

It is important to elaborate on this idea. The method used to describe the structure of science used by Chen, and used in this first study, is co-citation analysis. Co-citation analysis is used to identify citation networks, or the islands that are populated with references. There are, however, subtle differences in how these co-citation models are created that influenced the design of this experiment. Chen uses a local model (papers and their references are retrieved based on topic search) and aggregates data over multiple years (Chen, 2012). By contrast, we use a global model (all of the documents in the Scopus database), create models each year to represent the socio-cognitive structure for that year, and then link models year-to-year to show changes in socio-cognitive structures. This approach, along with measurements of its accuracy, is described elsewhere (Boyack & Klavans, 2013; Klavans & Boyack, 2011). In addition, the a priori identification of all of the islands each year allows us to simplify the indicators. As mentioned above, pairs of reference can "vote" for innovativeness, conformity, or can be undecided.

Use of our global model also allows us to provide a relatively simple indicator of when the flow of science has been disrupted. One simply locates the island

associated with a high impact paper and looks at year-to-year discontinuity – when does the island appear and when does it disappear. It is this unique characteristic of our global co-citation model – the cluster start and end dates – that we are building on. As the dependent variable, we focus on the stability of the island (the citation network) most associated with a high impact paper. If a paper is a turning point paper, then the island associated with the turning point paper should be unstable or have increasing instability.

It is also important to note that, at this stage of the experiment, we were only interested in an indicator of innovative papers. We expected that the dependent variable (the stability of the island) would be negatively associated with innovativeness votes, positively associated with conformity votes, and not related to the undecided votes. Or, stated differently, we didn't know which measure would be better: innovativeness or not being conforming. Since we expected a large percentage of the votes would be undecided, it wasn't clear which indicator would by better.

The data used in this experiment consisted of over four thousand papers where we have full text data. This sample was based on first identifying the 16,427 top 1% (most highly cited by 2010 by discipline) papers that were published in 2007. A total of 4216 of these papers were found in the Elsevier 2007 full text corpus. Several of these papers did not have a sufficient number of references to generate a meaningful statistic. A total of 4180 papers had a minimum of 5 references and were also found in our 2007 co-citation model.

Table 1. Correlation of votes for innovation and conformity with co-citation cluster age (stability) using different distances between reference pairs.

Distance between pairs (in characters)	*0*	*375*	*1500*	*6000*	*ALL*
General Statistics					
Avg. #votes/paper	108	200	465	1126	4847
#papers (#votes $\geq$10)	3295	3992	4119	4156	4180
% Innovative votes	32.4	36.1	39.8	42.1	43.4
% Conforming votes	26.7	21.5	16.4	12.8	10.0
% Undecided votes	40.8	42.4	43.8	45.1	46.6
Pearson correlation with stability					
% Innovative votes	-0.307	-0.263	-0.207	-0.163	-0.133
% Conforming votes	0.362	0.404	0.416	0.401	0.379
% Undecided votes	-0.039	-0.061	-0.064	-0.061	-0.052

Both full text and bibliographic data from Scopus were used to determine reference pairs. As shown in Table 1, varying distances between reference pairs were used. These distances corresponded roughly to same citing location (distance=0), same sentence (375 characters), same paragraph (1500 characters)

or same section (6000 characters). We also created the traditional co-citation pairs (all possible pairs in the bibliography).

The average number of votes per paper increases dramatically with distance between reference pairs. The average paper has only 108 votes (pairs of references) when one only count pairs of references in the same bracket. This increases to 1126 votes when one uses the largest distance (references roughly in the same section of a paper) and 4847 votes when one assumes that all references are related to each other once (the final column in Table 1). The number of papers that have meaningful indicators increases with distance between pairs. For example, if we assume that a paper had to have a minimum of 10 votes for an indicator to be meaningful, we could only create indicators for only 3295 (out of a possible 4180) papers if we use the smallest distance between pairs. Pairs of references are more similar (the % Conformity statistic goes up) as the distance between the pairs goes down. This is consistent with our recent findings proximate pairs of references are more similar than pairs of references that are far apart in the text (Boyack et al., 2013).

The drop in the sample size affects the correlation between the three dependent variables (%innovative; %conforming; %undecided) and stability (the total age of the thread is our surrogate for overall island stability). When we only use pairs that are in the same bracket, the indicator for innovativeness and conformity are highly correlated with stability and in the expected direction. But, as the distance increases, the correlation between the innovativeness indicator and stability deteriorates. By contrast, the correlation between the conformity indicator and stability increases slightly and remains surprisingly strong.

The fact that the innovative indicator and conforming indicator were not moving together was due to the increasing role of the undecided votes. When the distance between pairs was smallest, the undecided votes accounted for 40.8% of all votes and the correlation between %Innovative and %Conformity was relatively small (-0.20). As the distance increases, the percentage of undecided votes increases (to a maximum of 46.6%) and the correlation between innovativeness and conformity decreases (to -0.08). Basically, the indicator of conformity and innovative are weakly related to each other because of the large percentage of undecided votes. They act more independently of each other as the conditions change.

A regression model for these data (see Table 2, below) shows a similar pattern in the deteriorating impact of the innovative indicator and the strong impact of the conformity indicator.

In this case, we tested whether the number of years that an island survived after 2007 was a function of the age of the island in 2007 (expected to be positive), innovativeness (expected to be negative) and conformity (expected to be

positive). The T-statistics for each of these variables in the regression equation are all significant at the .001 level. All of the signs are in the expected direction. But most importantly, the pattern in the T statistics is consistent with what we observed in Table 1. The effect of the innovativeness indicator becomes weaker as the distance between pairs increases and the number of observations increases. The T statistic for the conformity indicator is high and remains strong over the entire domain.

Table 2. Regression statistics for the experiment of Table 1 where survival = f(Age, %Innovative, %Conforming).

Distance between pairs (in characters)	*0*	*375*	*1500*	*6000*	*ALL*
Avg. #votes/paper	108	200	465	1126	4847
#papers (#votes $\geq$10)	3295	3992	4119	4156	4180
T-statistic (age)	33.01	35.90	36.41	37.58	38.81
T-statistic (%Innovative)	-7.82	-7.20	-6.15	-4.43	-3.66
T-statistic (%Conforming)	10.41	11.79	13.02	12.53	11.93
Adjusted R-square	0.369	0.366	0.364	0.358	0.355

It is at this point that we realized that Chen's notion of betweenness and its implications worked, but in an unexpected direction. Instead of creating an indicator of innovativeness, the model has created a very strong indicator of conformity. And it was at this point that we realized the possible implications of the results. An indicator of conformity represented a significant research opportunity; we simply had never heard of any researcher developing conformity indicators. But more importantly, we realized that some very high impact papers were doing exactly the opposite of what we had assumed. They were re-enforcing existing beliefs rather than challenging them. We therefore designed a follow-up experiment to explore these issues further.

Follow-up experiment

Given the unexpected results of the preliminary experiment, and the potentially controversial nature of the implications of those results, we determined to conduct a more detailed experiment whose purpose was to test the robustness of the preliminary experiment. The central feature of this follow-up experiment was a new global model of science that had been recently developed by Waltman & van Eck (2012). Their model was unique in two respects. First, they clustered ten years of publication data in one pass (almost 10 million articles from the Web of Science). This, in itself, was a significant accomplishment. Second, this was the first time a global model was created using direct citation analysis and ten years of data. Further elaboration of their accomplishment is needed to appreciate its importance and corresponding application for developing an indicator of conformity.

Direct citation analysis (i.e. who cites whom) was the method that was originally used by Garfield (1973). But direct citation analysis had, for 50 years, only been used for local models (i.e. creating a historiography around key papers). Boyack & Klavans (2010) had created the first large-scale direct citation model using five years of Scopus and Medline data, but found that direct citation produced significantly inferior document clusters than co-citation analysis. Waltman & van Eck argued that a global direct citation model needed a much longer time window to be accurate – five years was too short a time to allow the direct citation networks to emerge. Their results, using ten years of data, were creating very meaningful results. After seeing their article posted on ArXiv in March, 2012, we contacted the authors immediately, and were impressed with the case examples that they provided to us. We started using their code in April, 2012 to replicate their results using the Scopus database. Overall, we found that Waltman & van Eck were correct – direct citation models based on a time period of 10 years or longer are just as accurate as co-citation models as measured using textual coherence (Boyack & Klavans, 2013). We therefore proceeded to push the boundaries further – creating a 16-year model (the longest period possible with the Scopus database).

A 16-year direct citation model and a 16-year co-citation model were developed over the next few months, and are both used in the follow-up experiment. Each of these global models is used to generate clusters of documents (islands). But the nature of the islands is quite different. The islands from a co-citation model represent a snapshot of the socio-cognitive belief systems for a single year. These belief systems are relatively unstable over time; many islands are sinking and many new islands appear each year. By contrast, the islands from a direct citation model represent a retrospective view (from 2011) about citation history. These islands are much more stable over time; relatively few islands sink or are born. A description of the methodology for created these two models can be found in Boyack & Klavans (2013).

We also changed the way that pairs of references are identified and how the votes are weighted in the second experiment. These are minor changes in methodology, and were done to explore the robustness of our findings using slightly different procedures. Specifically, we decided to use citances (sentences that includes references) instead of distance between pairs for two reasons. First, citances will allow us, in future studies, to look more deeply into the words that are used when one cites a paper. We have recently been exploring citance analysis as an alternative method for identifying breakthrough papers (Small, Boyack, & Klavans, 2013). There was no a priori reason why this approach might also be useful for identifying conforming papers. The second reason for citance analysis was to provide an alternative way to weight the votes. In the first experiment, all pairs of references were weighted equally. In this study, each citance gets one vote. Fractionalization was done based on the number of reference pairs in a

citance so that a very long citance with tens of references doesn't overwhelm the results.

The same data, indicators and dependent variables are used as in the first experiment. The correlations and regression analyses for the citance analysis were calculated for high impact papers with at least three citances (n=4025). The correlations and regression analysis for the bibliographic analysis were done for all papers with 10 or more votes and which could also be assigned to islands (4187 papers).

Table 3. Correlation and regression results from the second experiment.

Distance between pairs	*CC Citance*	*CC ALL*	*DC Citance*	*DC ALL*
Statistics				
#papers	4057	4187	4057	4187
% Innovative votes	22.1	29.8	37.8	44.4
% Conforming votes	20.9	7.3	18.7	5.5
% Undecided votes	57.0	63.0	43.6	50.2
Pearson correlation with stability				
%Innovative votes	-0.132	-0.008	-0.144	-0.029
%Conforming votes	0.345	0.362	0.368	0.370
%Undecided votes	-0.190	-0.176	-0.122	-0.115
Regression statistics				
T-statistic (age)	29.8	30.1	29.3	29.9
T-statistic (%Innovative)	-4.0	0.54*	-4.1	-1.6*
T-statistic (%Conforming)	16.3	18.1	15.4	14.3
Adjusted R-square	0.284	0.286	0.285	0.269

* not significant

These results are consistent with those reported in Tables 1 and 2. The indicator for conformity has a high correlation with stability, while the indicator for innovativeness has a poor correlation with stability. In the regression equations, the impact from the conformity indicator remains strong while the impact from the innovativeness indicator is weak or not significant. Overall, the correlations are slightly lower than those reported in Tables 1 and Table 2, which may easily be attributed to choosing citance analysis (which corresponds roughly to the 375 column in Tables 1 and 2) and the alternative weighting procedure.

Additional analyses were not considered necessary at this point. Our intent was to determine if the conformity indicator was robust using a different socio-cognitive model of science and a different method for identifying and weighting reference pairs. It was not our intent to figure out how to maximize the correlations or r-square values. And while we did explore different transforms (to deal with

skewness in the dependent and independent variables), the overall results remained the same. The correlation between the conformity indicator and the stability of the clusters associated with high impact papers remained very high (in some cases exceeding 0.40) and did not drop below 0.345. Different models and different weighting systems did not result in a deterioration of this relationship. Clusters are more likely to survive if they contain a high impact conforming paper, even after adjusting for the history of the cluster.

Discussion and implications

What started as an operationalization of Chen's computational theory of innovativeness has had unintended consequences: the development of a paper-level indicator of conformity (for high impact papers) that is relatively robust. One could use either a co-citation or a direct citation model to determine the percentage of reference pairs that are in the same cluster. The data from both models could be combined. The use of full text is not necessary; comparable results (in terms of explanatory value) can be found using the bibliography at the end of a paper and making the traditional assumption that all references are equally related to each other. There are, however, significant shortcomings in this study that should be emphasized at this point.

First, the choice of the dependent variable (i.e. the stability and survival of clusters) is not optimal. It was used as an indirect indicator, which is appropriate if one is using large sample size and scanning for useful indicators. A more direct indicator of conformity, such as author opinions of the innovativeness or conformity of their own papers, or a completely different research design is needed in order to proceed further.

Second, we have not taken into account the cluster associated with the citing paper. We are actually dealing with triplets – pairs of reference papers and the citing paper. The computational theory was not initially formulated in this way and may need to be revised. For example, if the citing paper is in one cluster, and both of the cited papers are together but in a different cluster, should this be a vote for conformity or innovativeness?

Third, more thought needs to be given to the large number of undecided votes. What do they represent? In the co-citation model, they are mostly references that have a low citation rate in 2007- they represent concept symbols that are not members of the socio-cognitive norms for 2007. But there is information embedded in these concept symbols. The question is- how can one pulls out this information? In the direct citation model, the undecided references are mostly older. There is no reason that these older references can't be assigned to reference clusters. Doing so would reduce the number of undecided votes.

Finally, thought needs to be given to the strategic implications of generating a paper-level indicator of conformity. There may be situations where high conformity is needed (i.e. helping to stabilize an exceptionally unstable environment). But there also may be situations where high conformity is attracting resources would be put to better use by funding innovative (and potentially risky) work. Further work is needed to unpack what is meant by conformity and the role (both positive and negative) it might serve in creating a vibrant and effective research system.

References

Boyack, K. W., & Klavans, R. (2010). Co-citation analysis, bibliographic coupling, and direct citation: Which citation approach represents the research front most accurately? *Journal of the American Society for Information Science and Technology, 61*(12), 2389-2404.

Boyack, K. W., & Klavans, R. (2013). Advances in constructing highly detailed, dynamic, global models and maps of science. *Journal of the American Society for Information Science and Technology, under review*.

Boyack, K. W., Small, H., & Klavans, R. (2013). Improving the accuracy of co-citation clustering using full text. *Journal of the American Society for Information Science and Technology, in press*.

Chen, C. (2012). Predictive effects of structural variation on citation counts. *Journal of the American Society for Information Science and Technology, 63*(3), 431-449.

Chen, C., Chen, Y., Horowitz, M., Hou, H., Liu, Z., & Pelligrino, D. (2009). Towards an explanatory and computational theory of scientific discovery. *Journal of Informetrics, 3*, 191-209.

Garfield, E. (1972). Citation analysis as a tool in journal evaluation. *Science, 178*(4060), 471-479.

Garfield, E. (1973). Historiographs, librarianship, and the history of science. In C. H. Rawski (Ed.), *Toward a theory of librarianship: Papers in honor of Jesse Hauk Shera* (pp. 380-402). Metuchen, NJ: Scarecrow Press.

Garfield, E. (1977). The 250 most-cited primary authors, 1961-1975. Part II. The correlation between citedness, Nobel prizes and Academy memberships. *Essays of an Information Scientist, 3*, 337-347.

Garfield, E. (2006). The history and meaning of the journal impact factor. *Journal of the American Medical Association, 295*(1), 90-93.

Garfield, E., & Malin, M. V. (1968). *Can Nobel Prize winners be predicted?* Paper presented at the 135th Annual Meeting of the AAAS.

Garfield, E., & Welljams-Dorof, A. (1992). Of Nobel class: A citation perspective on high impact research authors. *Theoretical Medicine, 13*(2), 117-135.

Geuna, A., & Martin, B. R. (2003). University research evaluation and funding: An international comparison. *Minerva, 41*, 277-304.

Hicks, D. (2012). Performance-based university research funding systems. *Research Policy, 41*(2), 251-261.

Klavans, R., & Boyack, K. W. (2011). Using global mapping to create more accurate document-level maps of research fields. *Journal of the American Society for Information Science and Technology, 62*(1), 1-18.

MacRoberts, M. H., & MacRoberts, B. R. (1989). Problems of citation analysis: A critical review. *Journal of the American Society for Information Science, 40*(5), 342-349.

MacRoberts, M. H., & MacRoberts, B. R. (1996). Problems of citation analysis. *Scientometrics, 36*(3), 435-444.

Martin, B. R. (2011). The Research Excellence Framework and the 'impact agenda': are we creating a Frankenstein monster? *Research Evaluation, 20*(3), 247-254.

Moravcsik, M. J., & Murugesan, P. (1975). Some results on the function and quality of citations. *Social Studies of Science, 5*(1), 86-92.

Nicholson, J. M., & Ioannidis, J. P. A. (2012). Conform and be funded. *Nature, 492*, 34-36.

Small, H., Boyack, K. W., & Klavans, R. (2013). *Identifying emerging topics by combining direct citation and co-citation*. Paper presented at the submitted to ISSI 2013.

Waltman, L., & Van Eck, N. J. (2012). A new methodology for constructing a publication-level classification system of science. *Journal of the American Society for Information Science and Technology, 63*(12), 2378-2392.

TRACING RESEARCH PATHS OF SCIENTISTS BY MEANS OF CITATIONS

Marianne Hörlesberger and Beatrix Wepner

marianne.hoerlesberger@ait.ac.at and beatrix.wepner@ait.ac.at
AIT Austrian Institute of Technology GmbH, Donau City Straße 1, A-1220 Vienna (Austria)

Abstract
This contribution presents an approach for tracing research paths of scientists based on their profile of cited references. The results could be applied for analysing the development of single or groups of scientists in their own area or into a new field. This might be interesting for evaluation for research grants or also for investigation of trans-disciplinary research centres. Our approach utilises the cosine powerfully and provides clear results.

Conference Topic
Management and Measurement of Bibliometric Data within Scientific Organizations (Topic 9); Science Policy and Research Evaluation: Quantitative and Qualitative Approaches (Topic 3)

Introduction
Has a scientist moved from one research field into another or has he / she stayed in the same research domain since the beginning of the research carrier? Such a question might be important for instance for supporting scientists to create their own research topics and environment, to give them the chance to build up an independent research field. A movement out of the familiar research area might be risky for a scientist, but would indicate an aspect of frontier research, which is an intrinsically risky endeavour [5]. Another application of our approach could be the answer to the question in trans- or interdisciplinary research institutes to make different approaches of scientists originating from various disciplines visible.

While most publications in bibliometrics, scientometrics, or informetrics deal with the analysis of a high number of scientists, or journals this contribution presents indicators for the investigation of individual authors. The shift of a scientist to a new domain can be unveiled by investigation of his / her cited reference profile.

The cited references of scientific publication represent the knowledge on which the research work is based on. Leaving the familiar research environment and moving into a new area entails changing the knowledge base of his / her research field [6].

The underlying hypothesis of our approach could be formulated as follows. If a scientist shifts to a new research domain he / she will cite different references to a greater extent than he / she has done in his / her previous work. Stepping out of the well-known knowledge base (expressed in the set of cited references) and research environment creates new opportunities as well as potential risk, and there is an interest in defining and identifying such a path and people changing from one field toward another [2]. These considerations lead us to develop indicators for such situations.

We consider the knowledge base on which the work of a scientist is built on.

Data

Since we closely work with a research centre "TRIBOLOGY", a very inter-disciplinary domain, where scientists from the various fields such as material science, engineering, physics, chemistry, or mathematics work together, we decided to investigate the development of the citation profile of 20 authors in this research area. These scientists with at least ten publications up to 200 over several years are taken into account.

The authors were selected from a general bibliometric study in "TRIBOLOGY" and combined with some authors from the research centre itself. For each author the publications available in Web of Science were recorded. The investigation in this contribution is therefore limited to Web of Science data.

Method

There are different possible concepts and indicators based on mathematics, bibliometrics, information theory, economic, or even physics which offer several ideas (as already described in many publication, under many others in (Boyack K W et al. 2010, [1]), (Chen C. 2005, [2]), (Czerwon HJ et al. 1995, [3]), Egghe L, 2009, [4]). Various disciplines provide algorithms and indicators for our research question, taking statistical, geometrical, or network analysis points of view into account. Moreover there are also approaches coming from economics (e.g. Gini coefficient), or of information theory (entropy) perspective. A discussion about different approaches is summarized in Leydesdorff and Rafols 2011 (in [7]), but they discuss the indicators in the context of "interdisciplinarity". The difference to our approach lies in the application to different datasets, thus we do not consider the interdisciplinarity but rather the specialisation and the development of individual authors in their own research field.

The concept idea of the introduced approach here is to apply the trigonometrically function cosine. But first the data has to be prepared appropriately.

We consider all publications of a scientist recorded from Web of Science[175]. Then we list all references and allocate them to all publications in the following way. If a considered cited reference (CR) is cited by paper *X*, the entity value is *1*. In case the considered CR is not cited by paper *X*, the entity value is *0*. This is done for all papers of a considered author.
In this way we get a vector e.g. *(1, 0, 0, 1, 1, 0, 0, 0, 1, ...)* for each paper. Then for each couple of papers, the cosine is calculated.

The trigonometric function cosine is a function of an angle. Is the angle orthogonal the cosine takes the value *0*. Is the angle *0* the cosine takes the value *1*. In other words the cosine takes the value *1* in case the two vectors are identical and takes the value *0* if the two vectors are orthogonal, because the inner product of two vectors (the numerator in F1) is *0*. The cosine of two vectors *a, b* is given by the following formula

$$\cos(a,b) = \frac{a \cdot b}{|a| \cdot |b|} \qquad \textbf{F 1}$$

Let us apply these considerations to our data "cited references". We except a cosine value closer to *1* in case the cited references concur in two considered papers. If the cited references do not concur the cosine would take a value closer to *0*. These considerations are motivated by (Czerwon HJ et al. 1995, [3]).
In this way we get *n-1* cosine values for each author, when his / her number of papers is *n*. Now the cosine value path can be validated by checking the keywords, titles, abstracts so that big changes (a cosine value close to 0) from one paper to the next or hardly any changes (a cosine value close to 1) can be explained.
For getting an overview of the shift in the reference profile over all papers all cosine values are summarised and divided through the whole number of considered papers of an author. This approach allows us to compare all our 20 authors in the dataset and provides an interesting result.

Results

Individual authors

The cosine of the two CR vectors describes in a simple value how similar two articles are. It makes obvious how long a scientist works on a specific topic, if he / she works on various topics, or when he / she moves to a new topic.

Figure 1shows a cited reference profile of Author_L. The values on the x-axis (the zeros) indicate a strong change in the research topics. Although we consider here only cited references this hypothesis could be validated by investigating the titles, keywords, and abstracts of the considered papers. In areas where the cosine is

[175] Web of Science provide the cited references in a relatively well standardised format.

high strong change in the topic represented in the titles, abstracts, and keywords can be noticed.

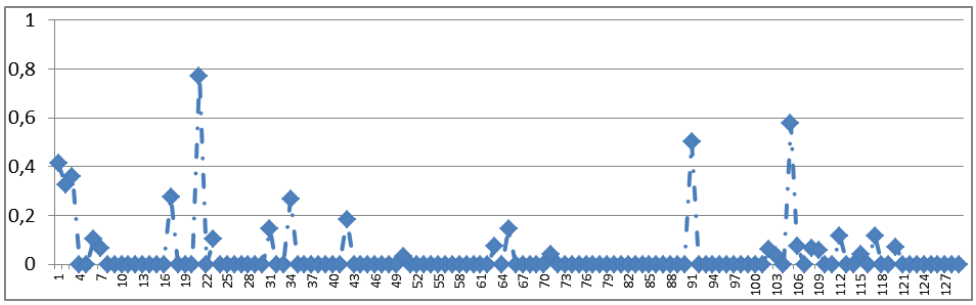


Figure 1. Cited reference profile of Author_L. The y-axis presents the cosine values. On the x-axis are the years (1 is the oldest, 127 is the youngest, whereas several papers are published in one year) of the papers publication.

A closer look on the papers uncover that the papers 16 and 17 for instance have very similar cited references. They deal with hip, knee replacement. The papers 52 to 61 cite different references, also 73 to 90, etc. Since paper 105 the term molybdenum disulfide occurs, which is discussed in the following publications. The last papers (121 to the end) are assigned to different topics in electrical wear, even though a similar field but considering various applications and materials thus reverting to different knowledge-bases.

The cited reference profile of Author_S is presented next (Figure 2). The research topics of this author have not changed so much resulting in a higher cosine value. This fact can be underlined by checking the keywords, abstracts, and titles of his / her papers. The research of Author_S develops more continuously with few interruptions, which can be seen in those periods where the cosine is zero.

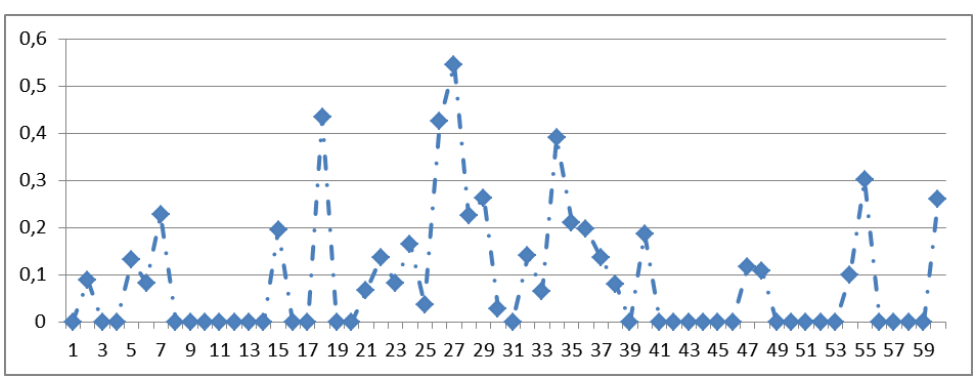


Figure 2. Cited reference profile of Author_S. The y-axis presents the cosine values. On the x-axis the papers are ranked in chronological order (1 is the oldest, 60 is the youngest, whereas several papers are published in one year).

The scientific career of the following Author_F (Figure 3) has developed continuously in a specific topic, namely physical vapour deposition, hard coatings, ultrathin coating, and nano (the high peaks in Figure 3). When the cosine values are zero the application field of his / her research has changed.

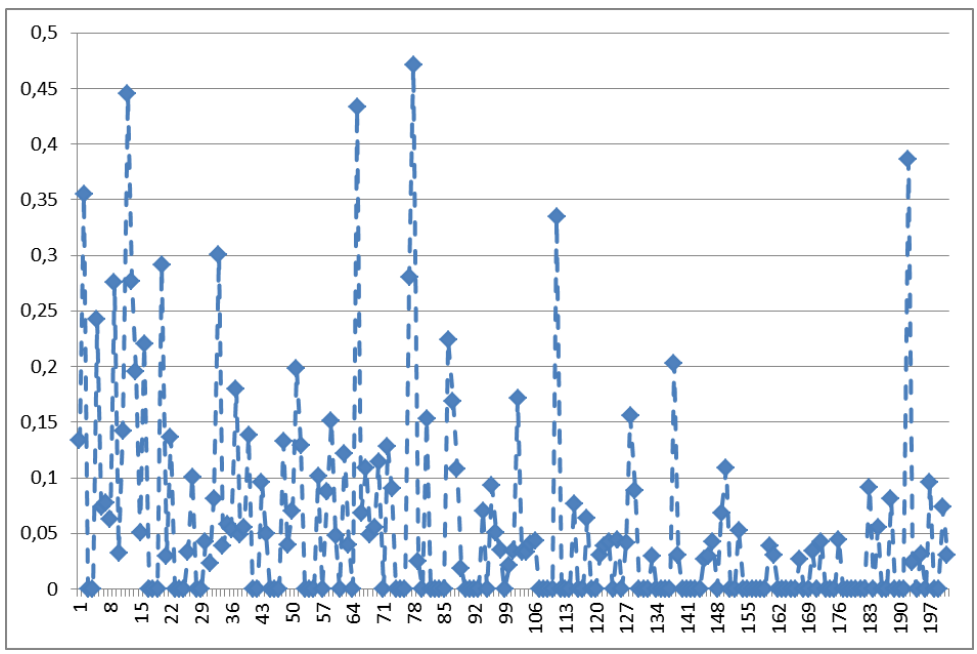


Figure 3. Cited reference profile of Author_F. The y-axis presents the cosine values. On the x-axis the papers are ranked in chronological order (1 is the oldest, with a long scientific career of 201 papers).

Author_T is a young scientist building up his research in one specific field. He /she has worked continuously on "turbulent boundary-layer flows". "Turbulences", "bluff body", or "separation" are some keywords describing his / her work. The cosine profile in Figure 4 demonstrates this fact clearly.

Author_O's cited reference profile indicates longer periods, where he / she works again and again on different topics. The peaks in the profile (Figure 5) show clearly the changes in the scientific works. They deal with "tribological properties", "nanoparticles", "lubrications", or "ultra-high-molecular-weight", "polyethylene/liquid crystalline polymer composites".

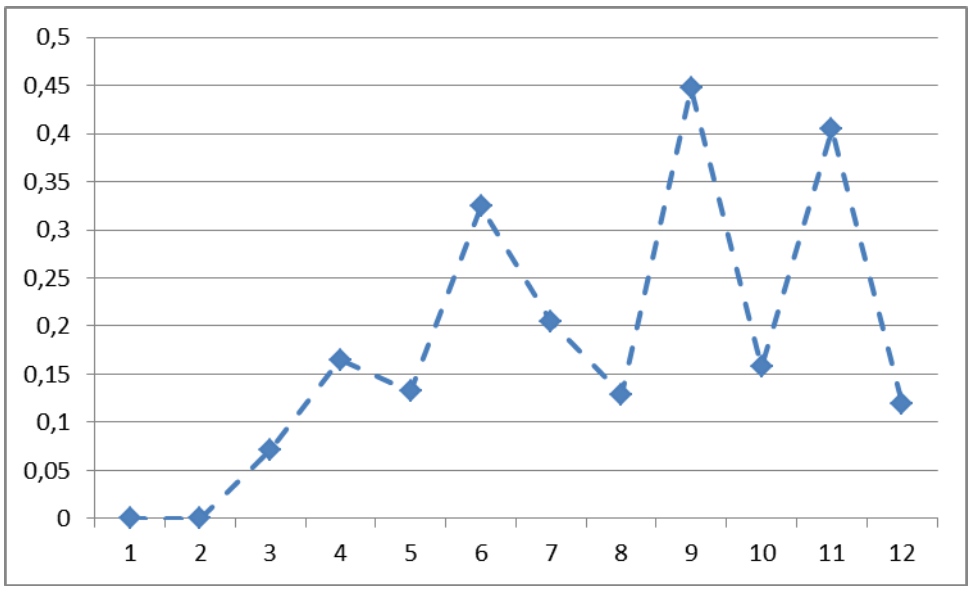


Figure 4. Cited reference profile of Author_T. The y-axis presents the cosine values. On the x-axis the papers are ranked in chronological order (1 is the oldest, 12 the youngest).

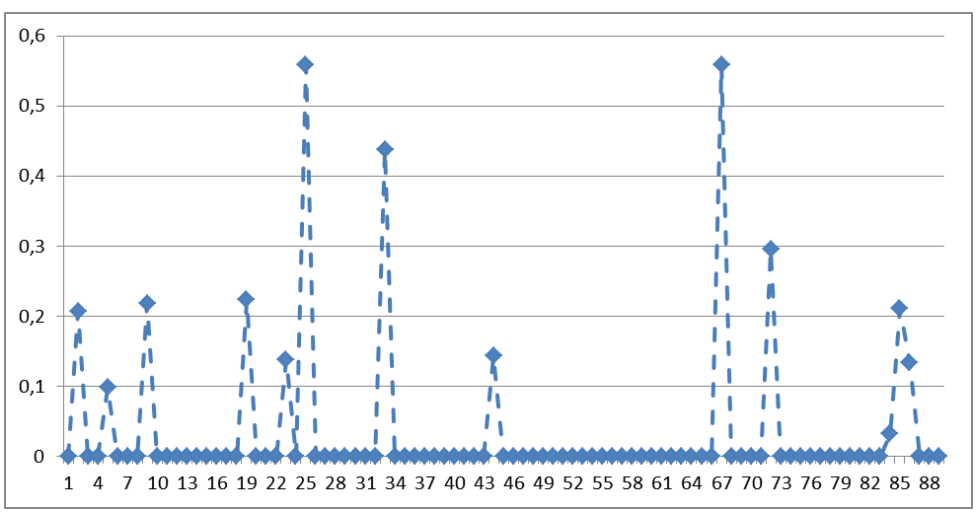


Figure 5. Cited reference profile of Author_O. The y-axis presents the cosine values. On the x-axis the papers are ranked in chronological order (1 is the oldest, 89 the youngest).

These samples confirm that the introduced approach is a useful method for investigating the shift to a new research field of a scientist.

A set of authors

Let us consider all investigated 20 authors and put their results into one chart. For doing so we applied the following procedure. All cosines of one author were summarised. After that this yield result was divided by the numbers of articles of each author. Then these values were ranked increasingly. The result is presented in Figure 6. When we summarise all cosine values of an author the amplitudes are weakened, of course. Nevertheless this one value (the sum of all cosine values of one author divided through the number of his / her publications) indicates generally if an author changed his / her research topic significantly or not.

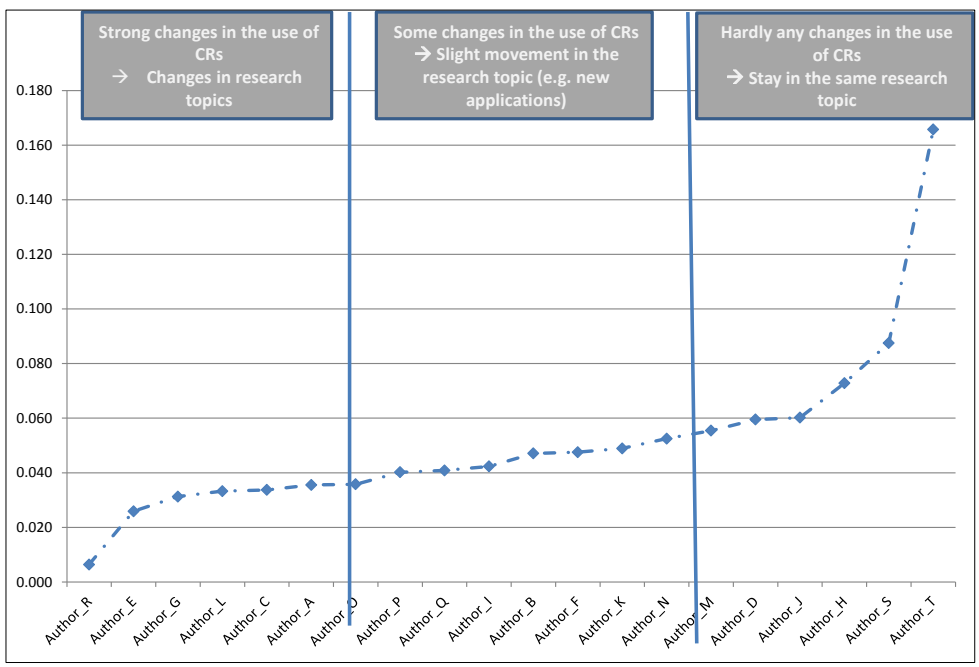


Figure 6. The overall cited reference profile of the 20 investigated authors.

Discussion

This approach unveils that if the articles placed in a chronological order the cosine clearly and easily shows differences in the knowledge-base a scientist reverts to. Also when comparing groups of authors using the cosine value summarised and standardized in relation to the number of papers of each author (Figure 6) provides a good insight to researcher publishing in different areas versus authors researching in constant and stable fields.

However the weakness of this linear approach is if an author published papers in two or more fields in parallel he appears to excursive. Therefore we work currently on developing a matrix method where all cited references of all papers

of an author are compared with each other so that parallel or intertwined research strings can be made visible.

References

[1] Boyack K W; Klavans R. (2010). *Co-citation analysis, bibliographic coupling, and direct citation: Which citation approach represents the research front most accurately?* Journal of the American Society for Information Science (JASIST). V 61(12), p 2389-2404.

[2] Chen C. (2005). *Measuring the movement of a research paradigm.* In Proc. of SPIE-IS&T, Visualization and Data Analysis. V 5669, p 63-76. San Jose.

[3] Czerwon HJ; Glänzel W. (1995). *A New Methodological Approach to Bibliographic Coupling and Its Application to Research-Front and Other Core Documents.* Proceedings of the 5th International Conference on Scientometrics and Informetrics, p 7-10. River Forrest, Illinois, US.

[4] Egghe L; Leydesdorff L. (2009). *The relation between Pearson's correlation coefficient r and Salton's cosine measure.* Journal of the American Society for Information Science (JASIST). V 60(5), p 1027-1036.

[5] Holste D; Scherngell T; Roche I; Hörlesberger M; Besagni D; Francoise C; Cuxac P; Schiebel E. (2012). *Capturing frontier research in grant proposals and initialanalysis of the comparison between model vs. peer review*. STI-Conference 2012, 5-8 September, Montreal.

[6] Hörlesberger M; Holste D; Schiebel E; Roche I; Francois C; Besagni D; Cuxac P. (2011). *Measuring the Preferences of the Scientific Orientation of Authors from their Profiles of Published References*. ENID (European Network of Indicator Designers) 7-11 September, Rome.

[7] Leydesdorff L; Rafols I. (2011). *Indicators of the interdisciplinarity of journals: Diversity, centrality, and citations.* Journal of Informetrics. V5, p 87-100. DOI: 10.1016/j.joi.2010.09.002

TRACKING ACADEMIC REGIONAL WORKFORCE RETENTION THROUGH AUTHOR AFFILIATION DATA

Sharon Kitt

s.kitt@ballarat.edu.au
University of Ballarat, University Drive, Mount Helen VIC 3353 Australia

Abstract
Academic mobility is considered a standard requirement for the development and progression of an academic research career. However, this career mobility is at odds with the drive to recruit and retain professionally-qualified workers in regional Australia, to ensure future generations of regional Australians have capacity to access higher education in their home region. To date, little work has been completed regarding the retention of active research staff in regional Australia. The purpose of this paper is twofold: to determine the viability of using author affiliation data as listed on publications to track an institutional cohort of authors by their affiliation; also, to determine if data analysed using this method revealed any insights regarding the retention of academic staff. Whilst using author affiliation data was found to be viable, it required extensive data manipulation and cleansing. Once analysed, the data revealed intriguing insights into the retention and movement of active academic researchers. Implications for regional higher education will be discussed.

Conference Topic
Science Policy and Research Evaluation: Quantitative and Qualitative approaches (Topic 3); Research Fronts and Emerging Issues (Topic 4); Management and Measurement of Bibliometric data within scientific organisations (Topic 9)

Introduction
Mobility in the academic research workforce has long been a desired goal and an oft-stated requirement of building a successful academic career, although this may not be entirely through choice but of necessity (Morano-Foadi, 2005). Harrigan (1997) cites three reasons that academic staff leave an institution: "involuntary (did not earn tenure, dismissed for cause); voluntary (dissatisfied with position, found better career opportunity, higher salary elsewhere); or end of career (retirement or death)." Research into academic workforce movement focuses on the voluntary component of this, usually related to international movements of staff (Cantwell, 2011; Hugo, 2005; Jöns, 2007) – that is, either emigration from or immigration to a defined country, on a fixed term or permanent basis. Such studies reveal the benefits to knowledge production of academic mobility, as "cultural and geographical proximity and distance continue to shape international

academic exchange and enable the identification of different national academic cultures around the globe" (Jöns, 2007, p102).

Consequently, many governments across the world are actively encouraging the trans-national movement of research staff through provision of information via online 'mobility portals' (http://www.mobility.org.au/outgoing/portals). Australia too has a program to encourage incoming and outgoing researcher migration (http://www.mobility.org.au/). Several reports commissioned by the Australian government explicitly detail this requirement for mobility in creating innovation (Cutler, 2008; DIISR, 2009). This globalisation of research effort and staffing (Fahey & Kenway, 2010) for Australian researchers has been hailed as a positive step forward for the creation of new paradigms of knowledge. It could be expected that large nation-states, such as the USA and Australia, would see a degree of internal movement within the higher education sector due to the same geographical proximity benefits, but there has been scant attention given to mapping the degree to which this may or may not occur.

Of course, in creating a positive environment encouraging global research staff movement, there is an associated impact on academic staff retention at the institutional level. Essentially turnover and retention are two sides of the same coin: turnover is the rate at which staff leave regardless of when they commenced; and retention is the rate at which employees from the same cohort stay at the institution (CIPD, 2012). Research into turnover focuses on the reasons employees leave, whereas retention research focuses on the reasons why they stay.

In a 2008 survey by the United Kingdom's Higher Education Funding Council for England, academic turnover rates were given as 6%, the lowest of any employment group within the higher education sector (UCEA, 2008), with other studies suggesting annual turnover rates of 12.8% (Glandon & Glandon, 2001).The concept of low academic turnover is more of a concern rather than high turnover, although both ends of the spectrum were identified as being problematic (UCEA, 2008) given the desire to have a mobile knowledge workforce. Turnover rates focus on academic staff as a homogenous group, without any connection between the turnover rate and the type of academic staff being lost to the institution. For example, would a turnover rate of 6% be acceptable if it contained the most-published research academics within an institution?

In terms of calculating retention rates, it is usually measured as the changes in workforce composition between two points in time, usually year-to-year. However, for examining academic workforce retention, and taking into consideration academic contracts and tenure, a more longitudinal perspective on retention is required. There is a paucity of data surrounding academic staff

retention rates, however two studies have found that retention over a 10 year period is around 50% (Harrigan, 1997; Kaminski & Geisler, 2012).

Although academic mobility is seen as important for researchers and knowledge creation, it is the antithesis of desire in establishing and retaining professionally-qualified individuals within regional settings. Regional Australia - much like regional areas in any large or sparsely populated country - faces significant challenges in not only attracting but retaining professionally-qualified staff such as medical workers (Humphreys, Jones, Jones, & Mara, 2002; Mills, Birks, & Hegney, 2010; Moore, Sutton, & Maybery, 2010), teachers (Plunkett & Dyson, 2011), engineers and accountants (Carson, Coe, Zander, & Garnett, 2010). Higher education of students in rural and regional areas requires an academic workforce to support this, which in turn requires the initial attraction and the longer term retention of these new professionals to the rural or regional area (Hugo & Morriss, 2010). However, recent reviews on academic workforce issues either barely mention rural and regional problems in passing (Hugo & Morriss, 2010) or ignore them entirely (Coates & Goedegebuure, 2012).

A further compounding issue for building a sustainable, regional academic workforce is that of the ageing of the Australian population, where even the academic workforce demographic is increasingly clustered towards retirement age (Hugo & Morriss, 2010). This is not a problem unique to Australia (Edwards & Smith, 2009); thus the pool of talent that can be drawn upon to fill positions vacated by retiring academic in regional Australia is diminishing world-wide.

Thus rural and regional Australia faces three strong challenges in developing and retaining an academic workforce: firstly, from the culture of academe itself that mandates mobility in determining academic career success; secondly from the challenges that face the retention of all professional groupings in a regional or rural environment; and thirdly, from an ageing and shrinking academic workforce.

Use of author affiliation data

Each time an author has an academic paper published, a related institutional address or institutional affiliation is also recorded. This institutional affiliation is used to identify the origin and further contact details of the authors using a variety of non-mandated, non-standardised address fields such as school, faculty, university name, campus name, suburb, state and country. This then forms part of basic citation data that can be exported electronically from bibliographic databases. For this study, the data were exported from the Elsevier citation database Scopus due to its availability in the author's home institution.

The use of institutional affiliation is not uncommon within the scientometric and bibliometric literature. These analyses are usually one of the following varieties:

- To compare a number of institutions by publication activity e.g. (Sorensen, 1994)
- To examine research activity by geographic areas, usually limited to a specific discipline e.g. in tropical medicine (Falagas, Karavasiou, & Bliziotis, 2006), in cardiovascular diseases (Rosmarakis et al., 2005), and in terrorism studies ((Reid & Chen, 2007)
- To undertake network analyses for research collaborations and their evaluations (Ye, Song, & Li, 2012)
- To determine journal quality by using institutional affiliation as a proxy (Agrawal, Agrawal, & Rungtusanatham, 2011; Gorman & Kanet, 2005).

Thus, institutional affiliation data is typically used as a representation of quality for a journal, an institution, a geographic area, or all three combined. What is not apparent in the literature is a way of using affiliation data to tackle questions surrounding the human resourcing issues of academic mobility via publication outputs.

Purpose

The aim of this paper is to firstly examine the feasibility of using author affiliation data as a method of tracking research workforce retention. Secondly, if feasible, to examine any insights the affiliation data may reveal regarding academic retention and movement. Thus degree to which the retention issues are currently at play within an Australian regional academic environment and how bibliometric data can assist in understanding the issue are the main foci of this paper

Method

A university based in regional Australia was chosen to be the case study for this research. Located over an hour from the state capital city, the institution itself has small but not an insignificant cohort of research staff occupying several nearby regional campuses. Importantly, the university itself had no metropolitan presence, making it feasible to examine the author affiliation data without fear of the data being biased due to the influence of urban campuses. The Scopus database was then interrogated to determine the name variations for the case-study university as due to institutional mergers and other reasons, an institution may have one or more names listed. Fortunately, in this case there was only one affiliation name listed. So, on the 25 September 2012 all journal papers published in the years 2005, 2006, 2010 and 2011 that had any authors listed with the case-study institutional affiliation were extracted from Scopus. These papers were extracted as two, two-year cohorts; 2005-6 forming the baseline cohort, with 2010-11 forming the second, comparison cohort. The two-year grouping of cohorts was selected for a variety of reasons: firstly, to account for disciplines that may have very different publishing frequencies; secondly, to account for staff that may have newly arrived at the institution and required some start up time; and

finally, it aligns with the formulae used by the Australian Government in calculating federal grant allocations, where the number of publications are averaged over a two year period. The five year period between cohorts was considered by the author to be adequate to allow for some indication of staff mobility to be evidenced.

From the affiliation search, 613 articles were extracted for the period 2005-6 and 2010-11. The bibliographic data extracted included author affiliations for all authors on the articles. An initial scan of the data resulted in one article being discarded due to the date of publication actually being 2012. Three other articles were discarded as they were still listed as being in press at 2011 so their official date of publication was unclear. One further article was discarded due to being a duplicate record in Scopus (confirmed and later deleted by Scopus). Out of an initial extraction of publication affiliation data, a total of 607 documents remained viable for analysis and form the basis of this study.

Data issues

Previous research using author affiliation data noted the requirement for extensive data cleansing and matching to ensure the accuracy of affiliation data (Neuhaus & Daniel, 2008), which was also found to be an issue within this study. Whilst it is possible to view disambiguated author affiliation data within Scopus as part of the normal viewing of citations, as an exported dataset affiliations were provided as a heterogeneous string variable within the same field. For example, a preliminary investigation revealed that whilst author names were mostly recorded consistently, their affiliation declarations were vastly different across all institutions within the extracted dataset. This made it difficult to identify the number of institutional authors on a paper without manually reviewing the data. In this case study, affiliation was necessary to identify only authors that had an affiliation listed with the institution. However in examining the data the variations within the affiliation fields included:

- variations of the name or abbreviated name of the university (such as "Univ of X", "Uni of X", "X Uni", et al)
- variations in the address of the university (PO box, street address, no address given)
- variations in the name of the suburb, or no suburb listed
- variations in the postal code/zip code of the address (e.g. 1234, 1233, no postcode)
- Inclusion or exclusion of the School or Faculty name
- Inclusion or exclusion of the State
- Inclusion or exclusion of the Country (Australia)

The variety within the data sets as per the above make difficult the development of programmatic filters or algorithms to easily cleanse the data into a homogenous

set. It also complicates searches conducted within the dataset, as data elements that purportedly reveal affiliation may be overlooked or mistakenly included. For example, whilst the University itself contains a unique city name, searching for the name of the city also revealed authors that have published with a university-affiliated staff member but were themselves affiliated with other city institutions, such as the city library. Within the bounds of this study it was important to exclude such author affiliations to ensure the accuracy of the results.

A case study approach to this study proved to be useful, as it was feasible to manually examine each of the authors affiliated with the institution and check each of the papers within the extracted dataset. That is, namesakes and name variants were identified and reconciled against institutional staff lists and other publication data (where available). In this way, author details were normalised within the dataset to ensure that the two cohort datasets where internally valid. Thus each affiliated author was represented by only one name in each cohort and paired across cohorts. Non-affiliated authors – that is, those that were not affiliated with the case study institution in either cohort, were not disambiguated nor normalised and therefore the numbers of authors should be taken as indicative only.

Author identification aside, an unanticipated side issue was that of incorrect author matching by Scopus, where publications by authors with similar names were associated incorrectly with an affiliated author. This was identified for five institutionally-affiliated authors. Requests were submitted to and acted upon by Scopus for correction in three of these cases, with the remaining two being sufficiently complicated cases requiring further manual intervention by the author themselves (these two authors and their publications were 'cleaned' by the author of this analyses for the purposes of the case study only). Whilst five authors per 159 is a relatively small error rate (3%), it would be of interest to delve further within the Scopus database to determine if this incorrect citation attribution is representative of all author data within Scopus. If so, this has considerable implications for the accuracy of Scopus data for bibliometrics and other research work.

Results

There were 607 journal articles reviewed for this study. In the 2005-2006 baseline cohort, 203 valid papers were used, with the remaining 404 papers in the 2010-11 comparison cohort. Table 1 (below) describes the characteristics of the cohorts.

Comparisons between the two cohorts paint an interesting picture of activity during these two time periods. In 2005-6 there were 330 instances of the institutionally-affiliated authors on 203 papers, with the average number of institutional authors per paper being 1.6. In 2010-11, the number of affiliated authors nearly doubled to 650; however, the average number remained the same

at 1.6 affiliated authors per paper. Similarly, there was a decrease in the relative percentage of unique lead authors, from 0.8 to 0.6 per paper. This means whilst the number of papers is increasing, this may be due to a higher publishing rate from a smaller group of authors, but less of these authors are actually named as first author. Whilst the average numbers of all authors per paper increased from 2.65 to 4.00, this figure was skewed by one paper that had more than 350 collaborators. Removing this paper as an outlier would reduce the average numbers of authors per paper for 2010-11 to 3.13.

Table 1. Scopus dataset characteristics

Description	*Baseline Cohort 2005-06*	*Comparison Cohort 2010-11*
No. of papers from Scopus used	203	404
No. of affiliated (normalised) and non-affiliated (non-normalised) authors on all papers	538	1614
No. of affiliated authors on all papers	330	650
No. of unique affiliated authors on all papers	159	241
Avg no. of affiliated authors per paper	1.6	1.6
Avg no. of unique affiliated authors per paper	0.8	0.6

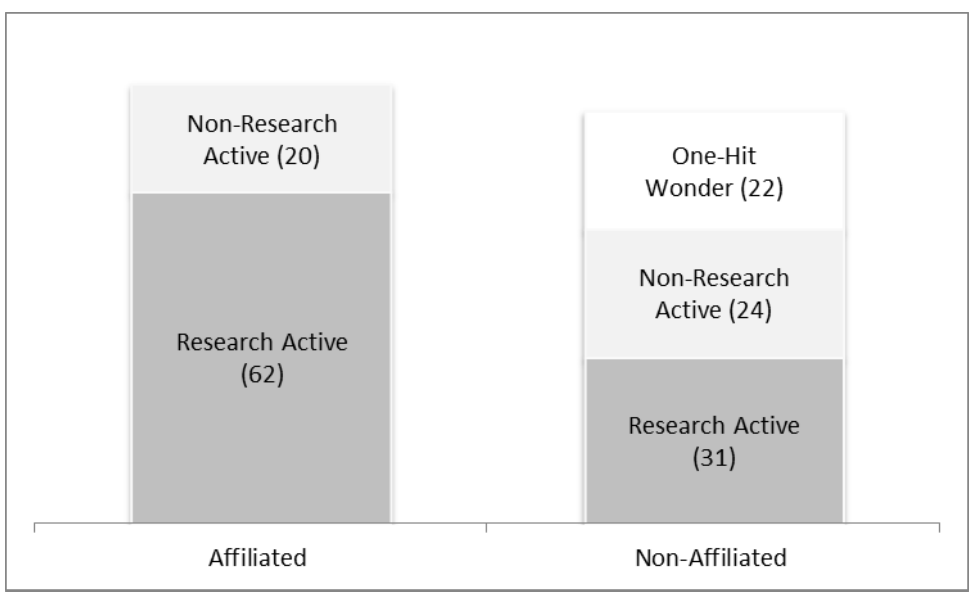


Figure 1. 2010-2011 publication activity of all institutionally-affiliated authors from the 2005-6 dataset

Looking at the differences between 2005-6 and 2010-11 (Figure 1), a picture of the retention activity of the 2005-6 baseline cohort can be seen. In summary, of the 159 authors listed in 2005-2006 baseline cohort, 62 (39%) authors were still

publishing journal articles affiliated with the institution in 2010-11 and were classified as "affiliated research active." Additionally, it means that of the 241 unique affiliated authors listed for the Institution in 2010-11, only a quarter of these (the 62 authors) were staff at the institution five years previously.

Through examination of the institutional staff contact directory from the 2010-2011 years, it was determined that 20 (13%) of the 2005-6 cohort were still employed at the institution but were not actively publishing – the "affiliated non-research active". By inspecting their position title and corporate location, it was clear that some of these affiliated non-active academic staff had at some point moved into senior administrative or executive positions of the institution. However, a number had retained their academic position but had not continued to publish academically.

Therefore, 97 authors (61% of the initial cohort) in total had not published papers affiliated with the institution in the subsequent cohort years. Whilst as described above, 20 of these 97 were still at the case-study institution, further investigation regarding these remaining non-affiliated group 77 authors was then conducted.

Of the 31 researchers that remained research active at other institutions, 20 moved to another Australian institution and 11 were publishing at an overseas institution. These are the research-active, non-affiliated group. Of these, only 2 of the 20 researchers that remained in Australia, remained at a regional institution – the rest were publishing from a metropolitan institution. So of the initial cohort of 159 active affiliated researchers, 11% were no longer situated within regional Australia. Essentially, the data suggests that if a researcher left the case-study institution to remain an active researcher, it was to continue their research career at a non-regional organisation.

For the non-research active, non-affiliated group, these were still 24 authors that had left the case-study institution and moved to another institution. However within the confines of the second cohort timeline, these 24 authors did not publish within the 2010-11 period. This was determined through examination of their Scopus author history record, which records the latest affiliation of the author as per their most recent publication. The current status of these authors in terms of whether they were still employed within their latest institution was deemed beyond the scope of the current study.

For those that were not affiliated, 22 (14% of the initial cohort) of these were no longer employed by the institution but had no further publications or affiliations listed for them within the Scopus database with any other institution. These were the proverbial "'one-hit wonders'" (Harzing & Van der Wal, 2008), only publishing one paper under their name. That is, 14% of the actively publishing

workforce had effectively disappeared five years later, lost to the case-study institution and academia generally.

Thus through tracking author affiliations through citation datasets, the retention rate for actively publishing academic staff from 2005-2006 to 2010-2011 was 52% over 5 years. This compares favourably to the retention rates reported in the literature. However, if only actively publishing academics were included, that retention rate drops to 39%.

Discussion

The primary purpose of this case study was to examine the retention of a cohort of academic staff from a single regional institution, through the use of their author affiliation data. It was found that only 39% of staff were actively publishing with the same regional institution after five years. This is much lower than the reported retention rate of academic staff of 50% across a ten year period. The difference may be accounted for as that this study examined active, publishing researchers as opposed to all academic staff within the institution. It is highly likely that the loss of 61% of actively publishing– and one would assume actively researching – authors would have a deleterious effect on the ability of the case-study institution to nurture and grow research activity. There is however a paucity of research regarding the reasons for attrition and the resultant impacts, particular in the context of regional higher education. Indeed, it may be possible that this retention figure could be higher, lower or representative of institutions throughout the Australian higher education sector. Of similar concern to regional higher education research production was the 13% of authors located within the institution but not actively publishing. To support these researchers to research and publish would assist in the amelioration of issues arising from an apparently regional transient workforce. It would be worth further consideration as to the reasons why this is occurring and how institutions could support these academics to remain actively publishing researchers. Further investigation on the results and impact of this retention rate is needed to fully assess these issues.

One surprising result was the number of authors with a sole publication. These authors make up 14% of the original publishing cohort and are a curious anomaly. They clearly are individuals that are capable of undertaking an academic career evidenced through their ability to create academic publications, but through unknown circumstances have not. Questions arise whether their separation from the institution – and indeed academic life – was voluntary, involuntary or due to end-of-career issues. Perhaps a further option – end of academic career – could be put forward as a fourth reason for career mobility. Further research could investigate why these individuals did not pursue an academic career and whether that is an anomaly of this regional institution, all regional institutions or all institutions in general. Additionally, if these authors have moved into industry, it

would be worth examining ways in which institutions can keep these authors involved with academia to perhaps enhance university-industry collaborations.

For the researchers from the 2005-2006 cohort that did leave the institution and remained in Australia, all but two out of twenty remained in regional Australian institutions. With such a limited case study it is difficult to determine whether this is an anomaly of the dataset used in this case-study or a representation of 'brain drain' from regional Australia. However this case study does suggest that retention of research active academic staff within regional Australia requires further attention to determine who is leaving regional institutions and more importantly, why. For the case study institution at least, it should be a cause for concern.

The other purpose of this study was to determine the feasibility of using citation datasets containing author affiliations as a data source for higher education research. This study supported that it was a feasible method, which in turn provided insight regarding the retention issues facing regional institutions. However, the size of the dataset used in this study needs to be taken into consideration when preparing to replicate this type of analysis with larger and/or multiple institutions. A key factor in the success of using this dataset was the ability by the study's author to track down authors on a case-by-case basis. It was relatively simple to identify the whereabouts of 159 authors after a period of 5 years; to take this per capita approach with author datasets that are larger (more authors) or more complex (multiple sites, multiple institutions) requires a considerable investigative investment. In the future, it may be that projects such as ORCID (Open Researcher and Contributor ID) may provide a way of ensuring this data is consistently recorded for all researchers, particularly as Scopus supports this initiative (ORCID, 2012). Meanwhile, automated or algorithmic approaches must ensure that they can operate within the data inconsistencies identified before they are able to be fully employed in such research.

Conclusion

This case study analysis undertaken is a novel approach to the use of bibliometric data, expanding the role author affiliation beyond that of merely a base to contact authors into readily accessible data to inform university human resource planning. It appears that retention of actively researching academic staff from this regional institution is low compared to expected retention rates of all academic staff. Additionally, the case study uncovered the phenomena of 'one-hit wonders', who may be a source of untapped talent that have moved into industry that could be better affiliated with institutions. This case study provides the first look at the movement of academic staff from the institution, where it appears that those that do separate from the institution do so to be employed at a metropolitan university. A corollary to this study will be an inward migration study conducted to examine who published at the institution for the first time in 2010, and from where they

originated. Whilst this case study shows that using the data in this way is viable, to do so requires considerable data reconciliation and validation. Regardless, the level of brain drain from a regional Australian institution has been tracked for the first time, and the loss of these active researchers to the region has considerable impacts on the institution – and regional higher education in general, now and in the future.

Acknowledgments

The author acknowledges the assistance of the Collaborative Research Network, University of Ballarat.

References

Agrawal, V., Agrawal, V., & Rungtusanatham, M. (2011). Theoretical and Interpretation Challenges to Using the Author Affiliation Index Method to Rank Journals. *Production and Operations Management*, *20*(2), 280–300.

Cantwell, B. (2011). Transnational Mobility and International Academic Employment: Gatekeeping in an Academic Competition Arena. *Minerva*, *49*(4), 425–445.

Carson, D., Coe, K., Zander, K., & Garnett, S. (2010). Does the type of job matter?: Recruitment to Australia's Northern Territory. *Employee Relations*, *32*(2), 121–137.

Coates, H., & Goedegebuure, L. (2012). Recasting the academic workforce: why the attractiveness of the academic profession needs to be increased and eight possible strategies for how to go about this from an Australian perspective. *Higher Education*, *64*(6), 875–889.

Cutler, T. (2008). *Venturous Australia Report* (pp. 1–36). Melbourne, Australia: Cutler & Company Pty Ltd.

Department of Innovation, Industry, Science and Research (DIISR). (2009). *Powering Ideas: An innovation agenda for the 21st century* (pp. 1–76). Canberra ACT: Commonwealth of Australia.

Edwards, D., & Smith, T. F. (2009). Supply issues for science academics in Australia: now and in the future. *Higher Education*, *60*(1), 19–32.

Fahey, J., & Kenway, J. (2010). Moving ideas and mobile researchers: Australia in the global context. *The Australian Educational Researcher*, *37*(4), 103–114.

Falagas, M. E., Karavasiou, A. I., & Bliziotis, I. a. (2006). A bibliometric analysis of global trends of research productivity in tropical medicine. *Acta tropica*, *99*(2-3), 155–9. doi:10.1016/j.actatropica.2006.07.011

Glandon, S., & Glandon, T. (2001). Faculty turnover and salary compression in business schools: Comparing teaching and research missions. *Journal of Applied Business Research*, *17*(2), 33–40.

Gorman, M. F., & Kanet, J. J. (2005). Evaluating Operations Management?Related Journals via the Author Affiliation Index. *Manufacturing and Service Operations Management*, *7*(1), 3–19.

Harrigan, M. (1997). An Analysis of Faculty Turnover at UW-Madison. *Association of Insitutional Researchers of the Upper Midwest's Annual Meeting* (pp. 1–8). http://apir.wisc.edu/retirement/An_Analysis_of_Faculty_Turnover_at_UW-Madison_1997.pdf

Harzing, A.-W., & Van der Wal, R. (2008). Google Scholar as a new source for citation analysis. *Ethics in Science and Environmental Politics*, *8*, 61–73.

Hugo, G. (2005). Demographic Trends in Australia's Academic Workforce. *Journal of Higher Education Policy and Management*, *27*(3), 327–343.

Hugo, G., & Morriss, A. (2010). Investigating the ageing academic workforce: stocktake. Retrieved from http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.177.2863&rep=rep1&type=pdf

Humphreys, J. S., Jones, M. P., Jones, J. a, & Mara, P. R. (2002). Workforce retention in rural and remote Australia: determining the factors that influence length of practice. *The Medical journal of Australia*, *176*(10), 472–6.

Jöns, H. (2007). Transnational mobility and the spaces of knowledge production: a comparison of global patterns, motivations and collaborations in different academic fields. *Social Geography*, *2*(2), 97–114.

Kaminski, D., & Geisler, C. (2012). Survival analysis of faculty retention in science and engineering by gender. *Science*, *335*(17 February), 864–866.

Mills, J., Birks, M., & Hegney, D. (2010). The status of rural nursing in Australia: 12 years on. *Journal of the Royal College of Nursing Australia*, *17*(1), 30–37.

Moore, T., Sutton, K., & Maybery, D. (2010). Rural mental health workforce difficulties: a management perspective. *Rural and remote health*, *10*(3), 1519.

Morano-Foadi, S. (2005). Scientific Mobility, Career Progression, and Excellence in the European Research Area. *International Migration*, *43*(5), 133–162.

Neuhaus, C., & Daniel, H.-D. (2008). Data sources for performing citation analysis: an overview. *Journal of Documentation*, *64*(2), 193–210.

ORCID (2012). About us. Retrieved January 13, 2013, from http://about.orcid.org/

Plunkett, M., & Dyson, M. (2011). Becoming a Teacher and Staying One: Examining the Complex Ecologies Associated With Educating and Retaining New Teachers in Rural Australia. *Australian Journal of Teacher Education*, *36*(1).

Reid, E. F., & Chen, H. (2007). Mapping the contemporary terrorism research domain. *International Journal of Human-Computer Studies*, *65*(1), 42–56.

Rosmarakis, E. S., Vergidis, P. I., Soteriades, E. S., Paraschakis, K., Papastamataki, P. a, & Falagas, M. E. (2005). Estimates of global production in cardiovascular diseases research. *International journal of cardiology*, *100*(3), 443–9.

Sorensen, J. (1994). Scholarly productivity in criminal justice: Institutional affiliation of authors in the top ten criminal justice journals. *Journal of Criminal Justice*, *22*(6), 535–547.

The Chartered Institute of Personnel and Development (CIPD). (2012). Employee turnover and retention. *HR Factsheet*. Retrieved November 21, 2012, from http://www.cipd.co.uk/hr-resources/factsheets/employee-turnover-retention.aspx

UCEA Universities & Colleges Employers Assocation. (2008). *Recruitment and Retention of Staff in Higher Education 2008* (p. 20). Retrieved from http://www.universitiesuk.ac.uk/Publications/Documents/Recruitment_and_Retention_2008.pdf

Ye, Q., Song, H., & Li, T. (2012). Cross-institutional collaboration networks in tourism and hospitality research. *Tourism Management Perspectives*, *2-3*, 55–64.

TRENDS OF INTELLECTUAL AND COGNITIVE STRUCTURES OF STEM CELL RESEARCH: A STUDY OF BRAZILIAN SCIENTIFIC PUBLICATIONS

Raymundo das Neves Machado [1] & Jacqueline Leta [2]

[1] raymacha@ufba.br
Universidade Federal da Bahia (UFBA), Instituto de Ciência da Informação, Departamento de Processos e Fundamentos da Informação, Rua Basílio da Gama, s/n – Campus Universitário do Canela (Brazil).

[2] *jteta@biomet.ufrj.br*
Universidade Federal do Rio de Janeiro (UFRJ), Instituto de Bioquímica, Programa de Educação, Gestão e Difusão em Biociências, Prédio do CCS, Bloco B - sala 39, Cidade Universitária (Brazil).

Abstract
The present work maps the intellectual and cognitive structures of Brazilian research on stem cell in the period 1991 - 2010. Using the technique of author co-citation, we found that stem cell research in the country was marked by core authors from medical areas in the first decade and gradually they were outnumbered by specialists in the field. The technique of co-word analysis indicates that Brazilian research on stem cell initially had a more experimental and basic nature while, in more recent years, it also assumed an applied nature. This situation is accompanied by a notably increase in the number of Brazilian scientific publications and authors within this field. In the study period we also noted a series of initiatives in Brazil to stimulate stem cell research in the country, which increased the country's worldwide recognition. Our results suggest that Brazilian research on stem cell is in line with that in the context of global science.

Conference Topic
Visualisation and Science Mapping: Tools, Methods and Applications (Topic 8)

Introduction
During the last two decades, the number of Brazilian scientific publications catalogued in the mainstream databases has increased notably (Regalado, 2010). This has been a result of a combination of many internal and external factors, including the continuous investment of the public sector in qualifying human resources and improving infrastructure as well as the inclusion of dozens of new Brazilian titles in major scientific databases (Leta, 2011).
The Brazilian impressive growth in terms of scientific publications led the country to the 13th position in the world's ranking. However, a careful look at this growth reveals that Brazilian's main efforts are devoted to issues or research

themes of more international interest, especially in the fields of biomedicine, biology, agriculture and physics (Glanzel *et al*, 2006; Leta, 2011). Countries with such performance can be included in the bioenvironmental model (REIST-2, 1997), and, in the case of Brazil, this particular feature has historical roots related to the flourishing of science in the country.

The recent presence and success of Brazilian science in the fields of biology and biomedicine can be observed by its participation, for instance, in large-scale projects. Although Brazil did not take part as a member of the Human Genome Project consortium, the country has made some important contributions to this "big science" project not only by cloning genes of specific interest but mainly by heading the Human Cancer Genome Project. This project, sponsored by both the Sao Paulo State Research Foundation (FAPESP) and the Ludwig Institute for Cancer Research, had 29 laboratories engaged in sequencing more than one million of expressed sequence tags related to some human cancers (Kimura & Baia, 2002). Evidence of Brazilian presence in global science arena has been highlighted by one of the most prestigious scientific journals, Nature. In 2011, Brazilian biomedicine was featuring on its cover (Nature Medicine, 2011) and the country's potentialities in some scientific fields, especially in some hot topics, including stem cell, were emphasized by the journal.

Stem cells are cells found in all multicellular organisms (including plants) with the potential to divide and differentiate into other specialized cells as well as to self-renew (Maron-Gutierrez *et al*., 2009). The concept was proposed in the beginning of 20th century but it took some decades before scientists could really manipulate these cells. The year of 1981 was the landmark in the history of stem cell research: when scientists succeed in getting the first embryonic stem cells from a mammalian. From then on, knowledge on stem cell, especially on those related to mammalians, has been expanded at a quick pace due to its potential to be used in human beings.

The first Brazilian studies on stem cell are dated from the late 1980´s, when the country's first paper (Correa *et al*., 1987) was published. Brazil has been pioneering in stem cell research in Latin America and has developed many initiatives to stimulate collaborative projects in the area. As a result, the country is among the top five countries producing induced pluripotent stem cells. Among the achievements is the approval of a federal regulation that determines norms for manipulating and developing research on embryonic stem cell, in 2005 (Rehen & Paulsen, 2007). The coordination of the largest stem cell clinical trial in the world (a 1,200-person study) (Nature Medicine, 2011) and the foundation of a Brazilian Network on Cell Therapy in 2008 (composed by 52 laboratories) are other examples of the country recent efforts to develop competencies in this hot topic.

In this scenario, the interest and the potential uses of stem cell knowledge in global health have stimulated bibliometric studies to better understand the dynamics of its production through science communication. Ho *et al*. (2003) studied the scientific performance of South Korea, Singapore, Hong Kong and Taiwan on stem cell from 1981 to 2001. Li *et al*. (2009) investigated the world's

scientific output on stem cell from 1991 to 2006, focusing on its main trends and patterns. Zhao & Strotmann (2011a, 2011b) carried out a study based on author co-citation analysis to identify the intellectual structure of research on stem cell published in the period of 2004 – 2009. An & Wu (2011) applied the method of co-word analysis to analyze journal articles related to stem cell in the period 2001-2010. Also, Cantos-Mateos *et al*. (2012) investigated the Spanish scientific output in stem cell research published in the period 1997-2007.

Drawing upon the results of these studies, we made the following questions: do intellectual and cognitive structures change in time for a given research field? Do intellectual and cognitive structures change among countries with different cultures for a given research field? Using the analyses of author co-citation and co-word, we present some preliminary data from Brazilian scientific publications on stem cell. This pilot study is part of a larger project that aims mapping the main intellectual and cognitive structures of stem cell studies in the last decades and in some countries, all emerging economies, with different cultures: Brazil, China, India, South Africa and Russia.

Methodology

Brazilian publications on stem cell were retrieved on December 4th, 2012, directly from Web of Science (WoS) the ISI/Thomson Reuters' on line database that is available for most of Brazilian research institutions. [176]

Combining the terms "stem cell" or "stem cells" as main topic and "Brasil" or "Brazil" as address, a total of 1,607 Brazilian publications was found for the whole period covered by WoS. Nevertheless, as the early 1991s may be taken as the turning point for stem cell research, publications dated before 1991 were not considered in this study. All metadata available for the Brazilian publications (that is, publications with at least one Brazilian address) published from 1991 to 2010 were then selected and downloaded. In order to identify and map the intellectual and cognitive structures of Brazilian knowledge on stem cells, two main analyses were processed: author co-citation analysis (ACA) (White & Griffith, 1981) and co-word analysis (CWA) (Callon, 1991). According to the literature (ex.: Chen & Ibekwe-Sanjuan, 2010), ACA allows the identification of "intellectual structure of a scientific knowledge domain in terms of the groupings formed by accumulated cocitation trails in scientific literature". As for CWA analysis, it may help "to identify the relationships between ideas within the subject areas presented in these texts" (He, 1999, p. 134), to "analyze research trends and generate hypothesis and discover knowledge" (Jeong & Kim, 2010, p. 242) and "it is a powerful technique for discovering and describing the interactions between different fields in scientific research" (Muñoz-Leiva *et al*. 2011, online)

The co-word analysis was based on keywords listed in the publications. Misspelled words were corrected and synonyms were grouped into single terms. After this "cleaning" step, we analyzed the data on author co-citation and co-

[176] WoS can be accessed through: http://apps.webofknowledge.com/WOS_GeneralSearch_input.do?last_prod=WOS&SID=1D6MpmmCDcbeemialnG&product=WOS&highlighted_tab=WOS&search_mode=GeneralSearch

words, using Bibexcel software, version 2012-06-13, Microsoft Excel and VOSviewer, version 1.5.2.
As one of the goals of this study was to investigate time changes in cognitive and intellectual structure, all analyses were processed in 5-year periods, 1991-1995, 1996-2000, 2001-2005 and 2006-2010. As our main goal was to identify and map the pairs of co-occurrences, we opted to carry on a descriptive analysis for both ACA and CWA, instead of carrying a multivariate analysis, which is more frequent in co-occurrence studies.

Table 1. Characteristics of Brazilian publications on stem cell in each period.

Characteristic	*5-year Period*			
	1991-1995	*1996-2000*	*2001-2005*	*2006-2010*
Publications (a)	37	106	326	1,138
Only Articles (b)	33	83	226	772
Authors (c)	132	474	1,466	5,313
Cited References (d)	1,090	2,812	7,625	30,021
b/a	0.89	0.78	0.69	0.68
c/a	2.86	4.47	4.50	4.67
c/b	3.21	5.71	6.48	6.88
d/b	33.03	38.88	33.74	38.89
Main Research Areas*	Plant Sciences (7) Neurosciences & Neurology (5) Life Sciences & Biomedicine - Other Topics (4) Research & Experimental Medicine (3) Immunology (3)	Hematology (15) Biochemistry & Molecular Biology (12) Plant Sciences (11) Veterinary Sciences (10) Neurosciences & Neurology (10)	Hematology (40) Research & Experimental (26) Medicine (26) Oncology (26) Immunology (24)	Cell Biology (102) Hematology (88) Research & Experimental Medicine (83) Neurosciences & Neurology (68) Transplantation (60)

Main research areas refer to the WoS classification. Details are found at WoS homepage (see note 1). The number of articles for each main research area is shown in parenthesis.

Results

For an overview of Brazilian publications in each of the 5-year period, a brief descriptive analysis based on bibliometric variables is presented in Table 1.
Brazilian research on stem cell has grown substantially since the 1991s, which can be observed both by the total number of publications and by the number of authors in the byline of the articles. Concerning the number of total publications (including all original articles, meeting abstracts, reviews, proceedings paper, letters and editorial materials), a larger increase than that observed for original

articles was found. However, the ratio author per publication had a lower increase than that observed for author per article, suggesting that a larger number of Brazilian researchers may be involved in producing new knowledge on stem cell.
Considering the research areas, a shift is noted along the four periods studied. Nevertheless, some areas tend to be more frequent among the top five listed in Table 1, as follows: Hematology, Cell Biology, Research & Experimental Medicine, Neurosciences & Neurology and Immunology. This trend is in accordance to the results shown by Li *et al*. (2009), who analyzed the global publication on stem cell, from 1991 to 2006. After classifying publications in 167 main fields, the authors found that Hematology, Cell Biology and Immunology were ranked among the top five in terms of number of publications. In the authors' words "at the outset of the 21st century, increasing attention was paid to the field of cell biology, while the number of stem cell related articles in cell biology went beyond the oncology for the first time in the year of 2006" (Li *et al.*, 2009, p. 45). In a more recent study, Cantos-Mateos *et al*. (2012) have also identified Hematology as one of the main fields among Spanish scientific publications on stem cell.

The author co-citation analysis

The most cited authors in Brazilian publications on stem cells are listed in Table 2. Although some of the most cited names appear in more than one period (as it is case of G Paxinos and EC Perini), we can note that the list of most cited authors varies considerably in each period.

Table 2. The most cited authors in Brazilian publications on stem cell in each period.

5-year Period			
1991-1995	*1996-2000*	*2001-2005*	*2006-2010*
Amendt K	Akin DE	Paxinos G	Pittenger MF
Feldberg W	Wilson JR	Ljungman P	Zuk PA
Guertzen PG	Gluckman E	Charbord P	Orlic D
James EK	Johansen DA	Dexter TM	Johansen DA
Paxinos G	Metcalfe CR	Glucksbe H	Perin EC
	Paxinos G	Perin EC	Gronthos S
	Vansoest PJ	Strauer BE	Dominici M
			Caplan AI

It is noteworthy mentioning that this analysis considered the first author only. This is in accordance to Zhao & Strotmann (2001, p. 674) statement about the measures of first author in ACA "First-author counting tends to identify researchers who have conducted highly influential studies and emphasize a researcher's unique areas of study and most influential contributions".
The set of most cited authors constituted the main sources of ACA. For this reason, some of the most cited authors do also appear among the co-cited authors (Table 3), suggesting a positive relationship between higher number of citations

and co-citation. The set of most cited authors constituted the main sources of ACA. For this reason, some of the most cited authors do also appear among the co-cited authors (Table 3), suggesting a positive relationship between higher number of citations and co-citation.

Table 3 lists the pairs of co-cited authors with the highest frequency of co-citation in each period as well as the respective research areas where they have published, according to WoS classification. In the 1991-1995 period, the co-cited authors published in journals classified, for instance, in Neurosciences & Neurology, the area with the second highest number of publications in the period (position rankings appear in the parenthesis). In the period 1996-2000, the pair Akin and Wilson, for instance, has published in journals classified in Agriculture, the eighth area in terms of number of publications in this period. By the second period on, we note that research areas spread, which may be an evidence of the extension of stem cell research to other fields.

Table 3: Co-cited authors and their respective research areas in Brazilian publications on stem cell in each period.

5-year Period	Author 1	Author 2	Research Áreas
1991-1995	Feldberg W	Guertzen PG	Neurosciences & Neurology (2nd); Life Sciences & Biomedicine - Other Topics (3rd); Research & Experimental Medicine (4th)
	Amendt K	Guertzen PG	
	Amendt K	Feldberg W	
1996-2000	Akin DE	Wilson JR	Veterinary Sciences (4th); Agriculture (8th)
	Johansen D	Metcalfe CR	Plant Sciences (3rd); Veterinary Sciences (4th); Life Sciences & Biomedicine - Other Topics (7th); Agriculture (8th)
2001-2005	Perin EC	Strauer BE	Cardiovascular System & Cardiology (17th); Pathology (16th)
2006-2010	Orlic D	Perin EC	Cell Biology (1st); Research & Experimental Medicine (3rd); Transplantation (5th); Pharmacology & Pharmacy (6th); Cardiovascular System & Cardiology (18th)
	Gronthos S	Zuk PA	Cell Biology (1st); Hematology (2nd); Immunology (7th) ; Oncology (8th); Surgery (10th); Biotechnology & Applied Microbiology (12th); Veterinary Sciences (16th)

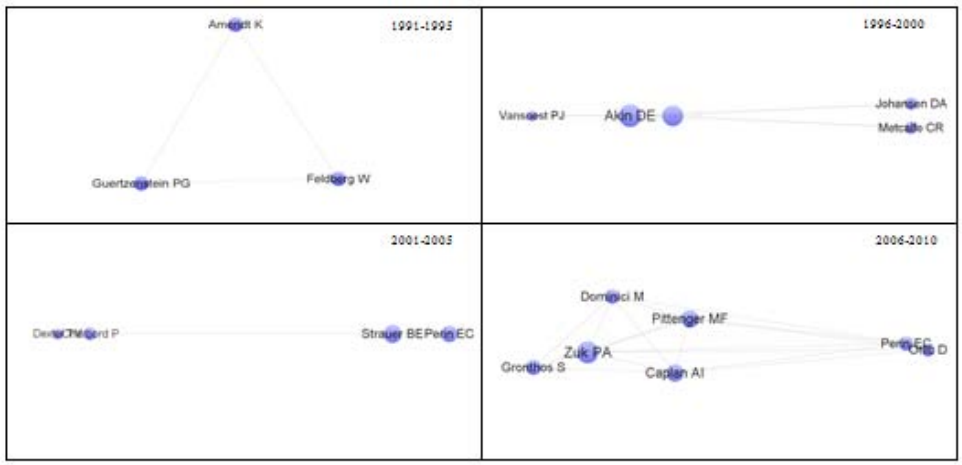
Research areas refer to the WoS classification. Details are found at WoS homepage (see note 1). The number of articles for each main research area is shown in parenthesis

Another relevant feature is the symmetric between the pairs of co-cited authors and the research areas. In the first period, the distribution of pairs and research areas indicates strong symmetry, *i.e*., there is a strong similarity between authors and research areas. This analysis shows three co-cited authors forming three pairs: (a) Feldberg and Guertzen, (b) Amendt and Guertzen and (c) Amendt and Feldberg. It is worth noting that the German Wilhelm Feldberg and the Brazilian Pedro Gaspar Guertzen were productive and widely known in the field of physiology; Amendt, also German, was an specialist in the field of cardiovascular research. Such trends explain the strong presence of the medical field in this period.

As for the second period, we found two pairs of authors highly co-cited: (a) Akin and Wilson and (b) Johansen and Metcalfe. Worth mentioning is that Danny E. Akin is a researcher at the Department of Agriculture, in the USA Agricultural Research Service; and Metcalfe is an specialist in the field of Botanics. For the third period, only one pair, Perin and Strauber, was high co-cited. Here, it is important to highlight that Strauber, a visiting professor at University of Rostock, in German, is one of the pioneers in the field of cardiac stem cell therapy. He is the author with both the highest frequency of co-citation and the highest density in Medicine, Transplantation and Cardiovascular System & Cardiology.In the fourth period of analysis, two pairs, (a) Orlic and Perin and (b) Zuk and Gronthos were the most co-cited authors. At least two of these authors are affiliated to a research institute devoted to the investigation of stem cell. Emerson C. Perin is the director of the Stem Cell Center at the Texas Heart Institute, in the US; Stan Gronthos, the co-director of the Centre for Stem Cell Research, Robinson Institute, at University of Adelaide, in Australia, who has patented a cell isolation technique to be used in stem cell preparations. Donald Orlic coordinates research projects in pluripotent hematopoietic stem cell, at the Hematopoiesis Section, from Genetics and Molecular Biology Branch, of the National Human Genome Research Institute/NIH.

Figure 1 shows the density of authors co-cited in Brazilian papers on stem cell, according to Van Eck & Waltman (2010). Authors shown in Figure 1 are also those with the highest levels of co-citation. As for the period 1991-1995, this set of authors shows the same level of density (as we may see by the size of the nodes), indicating that this network structure is well delineated. For the following periods, we note that authors differ in terms of density and that different clusters are formed, just as mentioned previously (Table 3).

Time trends for ACA suggest important changes in the intellectual structure of Brazilian publications on stem cell, mainly from the 2001-2005 period, when Brazilian research in this area seemed to take a new route, inspired by some key international specialists on stem cell.



Size of the nodes indicates the density of co-cited authors.

Figure 1: Author co-citation maps from Brazilian publications on stem cell in each period.

Co-word analysis

In order to map thematic trends in the field of stem cell, we carried out a co-word analysis (CWA) from keywords, as they "are the last link in the chain, used to shed light on the cognitive structure of a field […]" (Cantos-Mateos *et al*., 2012, p. 567). Along the four study periods, the number of keywords increased remarkably: from 191, in 1991-1995, to 3,246, in 2006-2010. As for the pairs of keywords, we observe another strong increase: from 4, in 1991-1995, to 29, in 2005-2010. In fact, a closer look at the data reveals that the share of keywords with frequency decreased, while the share of keywords with frequency higher than one increased (Table 4). This is an indicative that, over the periods, new terms were incorporated into stem cells' *corpus* of knowledge and constituted "hot keywords" for the field.

Table 4. Distribution of keywords' frequency of Brazilian publications on stem cell in each period.

5-year Period	Frequency of Keywords (with no duplication)						
	1	2	3	4-6	7-9	≥10	Total
1991-1995	174	12	4	1	-	-	191
(%)	(91.10)	(6.28)	(2.09)	(0.52)			(100)
1996-2000	379	43	10	10	1	-	443
(%)	(85.55)	(9.71)	(2.26)	(2.26)	(0.23)		(100)
2001-2005	975	134	33	40	11	10	1,203
(%)	(81.05)	(11.14)	(2.74)	(3.33)	(0.91)	(0.83)	(100)
2006-2010	2,420	433	144	143	37	72	3,246
(%)	(74.55)	(13.34)	(4.44)	(4.41)	(1.14)	(2.22)	(100)

Figure 2 illustrates three examples of keywords, whose frequencies increased during the four periods: bone-marrow, differentiation and transplantation. We note the same trend for these examples: keywords' frequencies were very low during the first two periods; in the following two periods, their frequencies increased n-fold and they became "hot keywords".

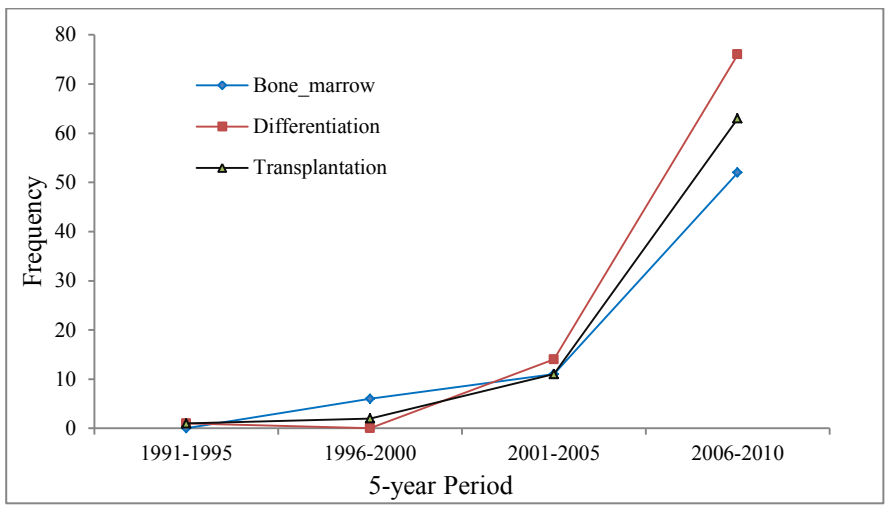


Figure 2: Time trends of three keywords extracted from Brazilian publications on stem cell in each period.

Hence, we can infer that these terms emerged in the beginning of 1991's and were not the focus of Brazilian research on stem cell at that time. In the last years, however, this picture changed and these terms (thematic) are now prominent.

Table 5 lists the keywords with highest frequency (in parenthesis) as well as the keywords with the highest number of co-occurrences. It is easy to note that some keywords appear with high frequencies in more than one period, especially after the third one. Worth mentioning is that this period is marked by the approval of a Brazilian regulation establishing rules for manipulating and developing research on embryonic stem cell. Also in this period, Brazil became one of the pioneers in using adult stem cells to treat some heart diseases and in establishing a novel procedure, a transplant of bone marrow cells in patients with heart failure due to Chagas' disease (Mendonça, online, p. 33).

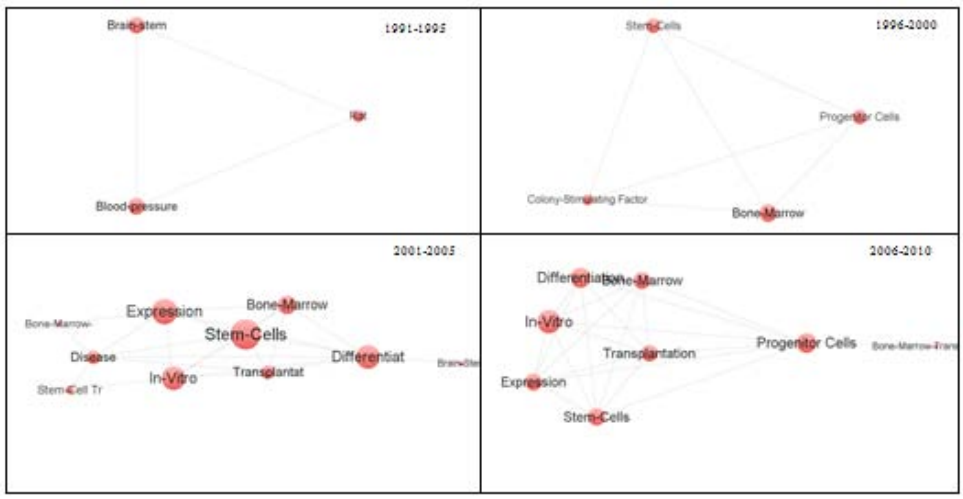
From the third period on, two keywords call our attention: differentiation and transplantation. They are on the basis of stem cell research. Differentiation is the process by which any cell (not specialized) becomes specialized; a basic characteristic of stem cells. On the other hand, transplantation is an application of stem cells. Another key term is bone marrow, which co-occurs with terms as in-vitro, expression, transplantation, progenitor cells and differentiation. Together, these terms not only point to the basic aspects of stem cell research but also underline its applied feature, which is clearly noted in the more recent periods.

Table 5: List of keywords with higher frequencies and with higher number of co-occurrences from Brazilian publications on stem cell in each period.

5-year Period	Keywords selected for CWA	Keywords with higher number of co-occurrences
1991-1995	Brain-stem (4) Blood-pressure (3) Hematopoietic stem-cells (3) Rat (3) Stem-cells (3)	Blood-pressure Brain-stem Rat
1996-2000	Stem-cells (7) Bone-marrow (6) Cells (5) Colony-stimulating factor (5) Expression (5) Progenitor cells (5)	Stem-cells Bone-marrow Colony-stimulating factor Progenitor cells
2001-2005	Stem-cells (24) In-vitro (20) Expression (19) Brain-stem (16) Differentiation (14) Stem-cell transplantation (14) Disease (11) Bone-marrow-transplantation (11) Transplantation (11) Bone-marrow (11)	Differentiation Disease Expression In-vitro Stem-cells Transplantation Bone-marrow Bone-marrow-transplantation Stem-cell transplantation
2006-2010	Stem-cells (105) In-vitro (90) Differentiation (76) Progenitor cells (72) Expression (64) Transplantation (63) Bone-marrow-transplantation (60) Bone-marrow (52)	Progenitor cells Bone-marrow Differentiation In-vitro Stem-cells Transplantation Expression

Note: Numbers in parenthesis represent the frequency of each keyword in the respective period.

Figure 3 presents the map of keywords with the highest frequencies of density, indicating the time trends for themes in Brazilian publication on stem cell. The maps show that the number of keyword pairs increased from the first to the last period. Some of the keywords appeared in more the one period, while others disappeared or grew in importance in the field (larger nodes). In other words, some themes were giving way to others. In this cognitive movement, stem cell research starts with a more experimental and basic scope, which turns gradually into applied.



(Size of the nodes indicates the density of co-words).

Figure 3: Keyword maps from Brazilian publications on stem cell in each period.

Some Remarks and Conclusion

Using the techniques of author co-citations and co-words, we examined Brazilian output on stem cell research over time. Our main goal was to map time trends of intellectual and cognitive structures of stem cell research in Brazil. Despite their potential use, it is worth highlighting that both ACA and WCA generate data that are proxies, and not the reality, of the dynamics of scientific work. Also important is the fact the methodological choices we have opted (for instance: a single informational source and a search strategy focused in a few number of keywords) determined the set of studied publications.

Despite these limitations, we believe that the collection of data presented in this paper provide a striking picture of how stem cell research has been carried on in Brazil. In this pilot study, we did not applied techniques of multivariate analyses, as one would expect. Instead, we opted to focus in co-citation frequency analyses and bibliometric maps. Hence, using descriptive analyses, we found that the onset of stem cell research in the country had a strong base of authors from medical areas. But, along the periods, these authors were outnumbered by experts in the field. It was also possible to identify that Brazilian research initially had a more experimental and basic nature, which gradually assumed an applied nature. This situation is accompanied by a notably increase in the number of Brazilian scientific publications and authors within this field.

Our results suggest that Brazil, an emerging country with no established scientific tradition, may already have the key ingredients to become an important player in stem cell research and develop a prominent role in global science at large.

Research on stem cell is very challenging, especially for its many potential applications to the healing of many human illnesses. This aspect justifies the huge rush for new knowledge and new procedures in this field, which has involved an

"army" of researchers around the world and from different specialties. Therefore, studying the dynamics of author co-citations and co-words in this field may help to better understand how the intellectual and cognitive structure of a promising and highly dynamics field behaves. Hence, the next step of our work includes a comparative analysis between Brazil and other emerging countries.

Acknowledgments

We are grateful to CNPq (Conselho Nacional de Desenvolvimento Científico e Tecnológico) for financial support and to Sonia Maria Ramos Vasconcelos (IBqM – UFRJ) for comments and review of the manuscript.

References

An, X.Y. & Wu, Q.Q. (2011). Co-word analysis of the trends in stem cells field based on subject heading weighting. *Scientometrics*, 88, 133–144.

Nature Medicine (2011). Biomedicine in Brazil. *Nature Medicine*, 11(10), 1169.

Cantos-Mateos, G. et al. (2012). Stem cell research: bibliometric analysis of main research areas through KeyWords Plus. *Aslib Proceedings: New Information Perspectives*, 64 (6), 561-590.

Callon, M. et al. (1991). Co-word analysis as a tool for describing the network of interactions between basic and technological research—the case of polymer chemistry. *Scientometrics*, 22(1), 155–205.

Chen, C. & Ibekwe-Sanjuan, F. (2010). The structure and dynamics of cocitation clusters: a multiple-perspective cocitation analysis. *Journal of the American Society for Information Science and Technology*, 61(7):1381409.

Correa, J.B.C. (1987). Ester-type lignin-xylan linkages in the cell-wall of stems of mimosa-scabrella. *Anais da Academia Brasileira de Ciências*. 59(5), 179-184.

Glänzel, W. et al. (2006). Science in Brazil. Part 1: a macro-level comparative study. *Scientometrics*, 67(1), 67–86.

Ho, Y. S. et al. (2003). Assessing stem cell research productivity. *Scientometrics*, 57(3), 369-376.

Jeong, S. & Kim, H-G. (2010). Intellectual structure of biomedical informatics reflectedin scholarly events. *Scientometrics*, 85, 541–551

Kmura, E.T. & Baia, G.S**.** (2002). Rede ONSA e o Projeto Genoma Humano do Câncer: Contribuição ao Genoma Humano**.** *Arq Bras Endocrinol Metab*, 46(4), p. 325-329.

Leta, J. (2011). Growth of Brazilian Science: a real internalization or a matter of databases´ coverage?. In: 13th International Conference of the International Society for Scientometrics and Informetrics, 2011, Durban, Africa do Sul. Proceedings of the ISSI 2011 Conference. Leiden, Zululand: Leiden Univ & Zululand Univ, v. 1. p. 392-397.

Li, Ling-Li . et al. (2009). Global stem cell research trend: Bibliometric analysis as a tool for mapping of trends from 1991 to 2006. *Scientometrics,* 80 (1), 41–60.

Maron-Gutierrez, T. et al. (2009). Terapia com células-tronco na síndrome do desconforto respiratório agudo. *Rev. bras. ter. intensiva*, 21(1), 51-57.

Mendonça. V. L. Possibilidades de utilização de células-tronco humanas e os obstáculos a serem vencidos para viabilizar seu uso em terapia. Available: http://genoma.ib.usp.br/educacao/A_USP_vai_a_sua_Escola_parte4.pdf. Acesso em: 11 jan. 2013.

Muñoz-Leiva, F. et al. (2011). An application of co-word analysis and bibliometric maps for detecting the most highlighting themes in the consumer behaviour research from a longitudinal perspective. *Qual Quant,* Available: http://sci2s.ugr.es/publications/ficheros/QuQu.pdf., Acesso: 5 jan. 2013.

Regalado, A. (2010). Brazilian Science: Riding a Gusher. *Science*, 330 (6009), 1306-12.

Rehen, S. & Paulsen, B. (2007). Células-tronco: o que são? Para que servem? Rio de Janeiro: Vieira & Lente.

REIST-2. (1997). The European Report on Science and Technology Indicators 1997.EUR 17639. Brussels: European Commission.

Van Eck, N.J. & Waltman, L. (2009). How to normalize cooccurrence data? an analysis of some well-known similarity measures. *Journal of the American Society for Information Science and Technology*, *60*(8) , 1635–1651.

Van Eck, N.J. & Waltman, L. (2010). Software survey: VOSviewer, a computer program for bibliometric mapping. *Scientometrics*, 84(2), 523–538.

Zhao, D. & Strotmann, A. (2011). Intellectual structure of stem cell research: a comprehensive author co-citation analysis of a highly collaborative and multidisciplinary field. *Scientometrics,* 87, 115–131.

Zhao, D. & Strotmann, A**.** (2001). Counting first, last, or all authors in citation analysis: a comprehensive comparison in the highly collaborative stem cell research field. *Journal of the American Society forr Information Science and Technology*, 62(4), 654–676.

White, H.D. & Griffith, B.C. (1982). Authors as markers of intellectual space: Co-citation in studies of science, technology and society. *Journal of Documentation*, 38(4), 255–272.

USE OF ELECTRONIC JOURNALS IN UNIVERSITY LIBRARIES: AN ANALYSIS OF OBSOLESCENCE REGARDING CITATIONS AND ACCESS

Chizuko Takei[1], Fuyuki Yoshikane[2] and Hiroshi Itsumura[3]

[1] *naoe.chizuko@ynu.ac.jp*
University of Tsukuba, Graduate School of Library, Information and Media Studies, 1-2 Kasuga, Tsukuba, Ibaraki (Japan)

[2] *fuyuki@slis.tsukuba.ac.jp,* [3]*hits@slis.tsukuba.ac.jp*
University of Tsukuba, Faculty of Library, Information and Media Science, 1-2 Kasuga, Tsukuba, Ibaraki (Japan)

Abstract
This research analyzes the obsolescence of citations and access with respect to a broad range of subjects, including fields that have not been dealt with in previous research, shedding light on the differences between these two types of obsolescence and the characteristics for each field. The targets of analysis were 473 journals that were randomly sampled from all of Springer's 11 subject fields. Using metrics such as Cited Half-life and Download Half-life, the study investigated the relationship concerning the rate of obsolescence for citations and access. As a result, comparatively strong and significant correlations were seen in "Chemistry and Materials Science" and "Engineering" with respect to long term obsolescence, and "Biomedical and Life Sciences" with respect to short term obsolescence ($p < 0.05$).

Conference Topic
Management and Measurement of Bibliometric Data within Scientific Organizations (Topic 9) and Bibliometrics in Library and Information Science (Topic 10).

Introduction
Since the early 2000s, package-type contracts for electronic journals, the so-called Big Deal, have been rapidly adopted among university libraries in Japan. Irrespective of the university's size, the Big Deal drastically increased the number of journal titles that could be accessed at contract universities. It also played a part in bridging the digital divide between universities and in promoting the development of information infrastructures there. However, future prices and access to the Big Deal depend on historic expenditures of an institution as well as current cancellations. With on-going budget cuts and increasing prices of journals, price hikes for the Big Deal are putting pressure on library budgets. This situation makes it difficult for libraries not only to maintain existing subscriptions, but also to subscribe to new journals. As withdrawal from the Big Deal results in a

dramatic decrease in the number of titles that can be used, and in turn, a collapse of the library's academic information framework, the building collections of journal backfiles that, in some way, alleviate the impact of these losses is a matter of urgency.

The development collection of journal backfiles differs from that of current files, which have a strong tendency to become fixed owing to budgetary considerations. This is because many universities have library staff introduce journal backfiles by utilizing sources such as special proposals received from publishers just before the accounting period to make decisions and proposals. However, few universities in Japan have sought to implement a planned introduction of journal backfiles by scrutinizing the level of on-campus demand and the effectiveness of such an introduction. Instead, universities have a strong tendency to select and purchase it within the limits of the amount left in the budget.

Previous research has been conducted on effective methods for introducing electronic journals largely focused on the new introduction of current files, such as the identification of core journals or the rate at which they can be supplied. However, there has been very little research on the effective development of journal backfiles. Investigations into the development of journal backfiles require perspectives focusing on the documents that fall into disuse, i.e., obsolescence. Obsolescence in books is investigated on the basis of the number of times a book is used by lending year or accession year. In the case of journals, on the other hand, their obsolescence is based on citations and access of materials. Understanding the relationship between both types of obsolescence will make it possible to estimate the obsolescence of access based on information about the obsolescence of citations. Correspondence between both has already been examined for certain fields, such as chemistry, and for specific journals. However, the nature of documental use (citations and access) varies by field, and it is possible that trends in the differences between both also differ by field. Thus, this research employs several indices of obsolescence, some of which have not been adopted in previous studies, and analyzes obsolescence in access and citations for a wide range of subjects, including fields that have not been examined in previous studies. We shed light on the differences between both types of obsolescence and their characteristics in each field.

Related Research

To evaluate journal collections, many studies have been conducted on the relationship between citations and downloads (access). For instance, Duy and Vaughan (2006) analyzed local citation data and Impact Factor (IF) with journal usage in the fields of chemistry and biochemistry. For the former case good correlations were seen, whereas no significant correlation was observed between IF and journal usage. Further examples can be found in McDonald (2007), Bollen and Van de Sompel (2008), or Watson (2009). In particular, there are several

studies on obsolescence in access and citation of electronic journals. Nicholas et al. (2005) surveyed synchronous obsolescence of access. They reveal that over half of all use was accounted for by items published within the last 15 months. Moreover, some studies have analyzed the relationship between citations and access by calculating the density of citations and that of access and then comparing both (e.g., Moed, 2005; Kurtz et al., 2005; Brody et al., 2006; Chu & Krichel, 2007).

In recent years, Schloegl and Gorraiz (2010; 2011) have performed more multifaceted studies related to oncology and pharmacology by using such indices as IF, Immediacy Index (II), and Cited Half-life (CHL). While in the case of oncology journals, their results showed that the mean of Usage half-life is 1.7 years, and CHL is 5.6 years in 2006, similar results were found in the case of pharmacology journals. Furthermore, they computed that so-called "download citation half-life ratio" and found a medium sized correlation between CHL and Usage half-life in pharmacology ($r = 0.42$). Wan et al. (2010) examined the relationship between Download Immediacy Index (DII) and citation indicators using the Chinese full-text database CNKI. They found DII has a potential to be a predictor for other indices such as h-index. While a moderate correlation between DII and II was observed in the field of agriculture and forestry ($r = 0.57$), in the case of psychology a strong correlation was found ($r = 0.8$). However, these analyses have only been performed on selected fields and journals, such as organic chemistry, astronomy, and astrophysics.

Methodology

The survey employed in this research focuses on Yokohama National University (YNU), Japan. YNU consists of four undergraduate colleges (College of Education and Human Sciences, College of Economics, College of Business Administration, and College of Engineering Science) and five graduate schools (Graduate School of Education, Graduate School of International Social Sciences, Graduate School of Engineering, Graduate School of Environment and Information Sciences, and Graduate School of Urban Innovation). The university's membership comprises around 600 full-time teaching staff, around 10,000 students (around 2,600 graduate and 7,500 undergraduate students), and 300 full-time non-teaching staff. It is a medium-sized national university that does not have a medical school in Japan.

This research employed Springer's usage data from January 2010 to August 2012. Springer is one of the publishers with whom YNU has the Big Deal subscription and provides statistics on the use of full texts by publication year (COUNTER Journal Report 5). We randomly sampled 473 journals from each of the 11 subject fields covered by Springer. The procedure of sampling is as follows.

We identified a total of 1,492 titles, excluding journals whose full text had never been accessed in YNU from the 2,912 titles available for the two years and eight months from January 2010 to August 2012. These titles were arranged in descending order of access count for each field. Springer's COUNTER Journal Report 5 defines the number of downloads, the number of times used, or the number of times accessed as the number of times the full text of an article is used. As with many studies, we employ this definition, and refer it here as access count. In order to survey the degree to which the scale of access count influences its obsolescence, these titles were then separated into three layers according to the cumulative ratio of access count in the field, considering that the cover ratio of the access count is often used to examine target subscriptions to electronic journals in Japan. 15 journals were randomly sampled from each of the layers; for layers with fewer than 15 journals, the total number of journals was taken to be the target of research. Table 1 shows all subscription journals by field. Here, journals whose data could not be obtained from Journal Citation Reports (JCR) and those with an access count of 0 were excluded. Among these, 473 titles became the targets of research for this survey (the number of titles: 32 (BS) + 45 (BL) + 40 (BE) + 45 (CM) + 45 (CS) + 45 (EE) + 41 (EG) + 45 (HS) + 45 (MS) + 45 (MD) + 45 (PA) = 473).

Table 1. Number of titles for each layer of cumulative ratio of access count in each field.

Subject of Springer	*0%– less 70%*	*70%– less 90%*	*90%– 100%*	*Whole*
Behavioral Science (BS)	7	12	13	32
Biomedical and Life Sciences (BL)	31	56	123	210
Business and Economics (BE)	22	10	20	52
Chemistry and Materials Science (CM)	28	37	68	133
Computer Science (CS)	16	18	30	64
Earth and Environmental Science (EE)	22	30	55	107
Engineering (EG)	11	22	26	59
Humanities, Social Sciences and Law (HS)	18	19	17	54
Mathematics and Statistics (MS)	22	31	44	97
Medicine (MD)	40	47	51	138
Physics and Astronomy (PA)	22	23	35	80
Whole	239	305	482	1,026

In addition, the research was performed on the basis of the following hypotheses:

Hypothesis 1: Obsolescence is slow in fields such as humanities and mathematics for both citations and access, but is fast in fields such as physics and medicine.

Hypothesis 2: Indices that demonstrate a strong correlation differ by field.

In order to verify the above hypotheses, the situation in each field was examined using the following indices as measures of obsolescence:

(1) Obsolescence of citations:
    (1A) CHL, (1B) II/IF (ratio between II and IF)
(2) Obsolescence of access:
    (2A) DHL, (2B) DII/DIF (ratio between DII and DIF)

Each of these indices represents values as of 2011. Data for CHL, II, and IF were obtained from JCR. DHL refers to Download Half-life and was calculated in the same way as CHL. It signifies that "the number of journal publication years from the current year going back whose articles have accounted for 50% of the total downloads (access) received in a given year." DII and DIF correspondingly apply the definitions of II and IF to access and signify the access count generated during the period for calculating the corresponding index (DII as well as II: 2011; DIF as well as IF: 2009–2010). Note that these are used with the addition of one so as to avoid the division by zero. Because Springer's statistics contained sections in which access count for multiple publication years had been calculated together, this research employed access count divided by the number of years in the section as the access count for each year.

CHL and DHL express slower obsolescence as values become higher, whereas II/IF and DII/DIF express faster obsolescence as values become higher. In addition, while CHL and DHL are the indices of obsolescence of use that take into consideration long periods of time, II/IF and DII/DIF are indices that focus in particular on the change in usage during several years after publication. DII/DIF, the ratio between DII and DIF, has not been used until now in obsolescence analysis. However, given that use of journals is generally concentrated at the time directly after publication, it seems that DII/DIF would also prove useful as an index representing the characteristics of the nature of documental use. Therefore, it was used in combination with II/IF in this research. The survey examined the degree of accordance—that is to say, correlation—of obsolescence between citations and access for each field with respect to the long term (CHL and DHL) and the short term (II/IF and DII/DIF).

If the characteristics of each field are revealed like Hypothesis 1, it also becomes clear that the main target to develop journal backfiles is the field whose obsolescence is slow. In addition, if good correlations are found between the indices of citations and access in some fields by examining Hypothesis 2, the information of CHL or II/IF is enough to determine the strategy to collect journal backfiles for those fields. That is, it suggests the predictability of the use of journal backfiles by the information before introducing them in those fields. On

the other hand, as for the fields where no correlation is found between the two types of indices, we should take into consideration both types of indices to collect journal backfiles.

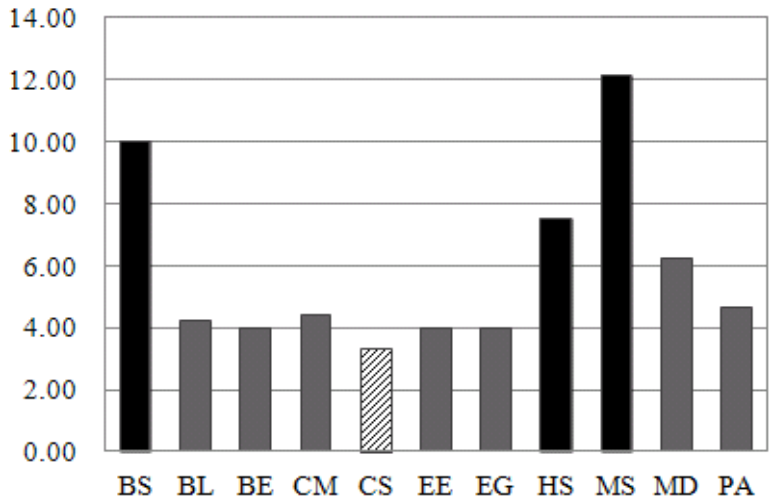


Figure 1. DHL by field.

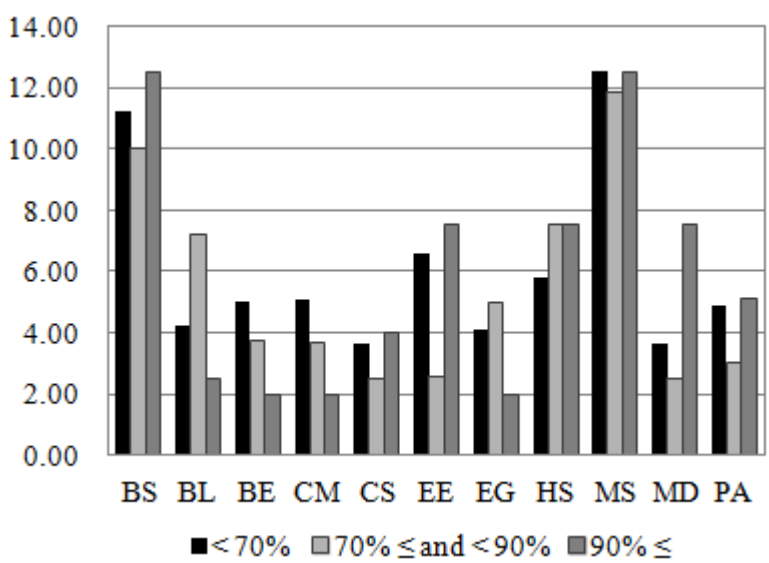


Figure 2. DHL by field and layer of cumulative ratio of access count.

Results and Discussion

General Situation regarding the Obsolescence of Access

Fig. 1 calculates the DHL of journals that were the target of analysis and sets out the resulting median values for each field. “Mathematics and Statistics” (12.14), “Behavioral Science” (10.00), and “Humanities, Social Sciences and Law” (7.50) are long, whereas “Computer Science” (3.33) is short. More detailed results are shown in Fig. 2, in which DHL values are calculated for each layer of the cumulative ratio of access count in each field. Whereas DHL values do not vary widely among the layers in the fields whose obsolescence is slow, the values tend to vary in the fields whose obsolescence is fast. DHL values show that the layer with the lowest access count (over 90%) tends to be shorter than other layers in the latter fields.

General Situation regarding the Obsolescence of Citations

Fig. 3 obtains the CHL of journals that were the target of analysis from JCR and totalizes the median values. “Mathematics and Statistics” had the largest value of CHL obtained from JCR, which is 10 (JCR describes values of CHL over 10 years as 10), whereas “Engineering” (5.40) and “Medicine” (5.70) had shorter values. Furthermore, Fig. 4 shows the content of Fig. 3 by layer of the cumulative ratio of access count. In contrast to DHL, no differences were seen by layer.

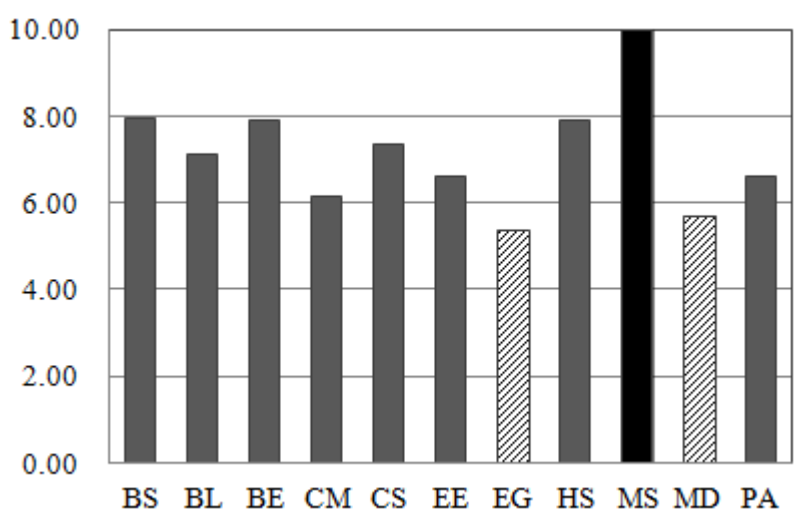


Figure 3. CHL by field.

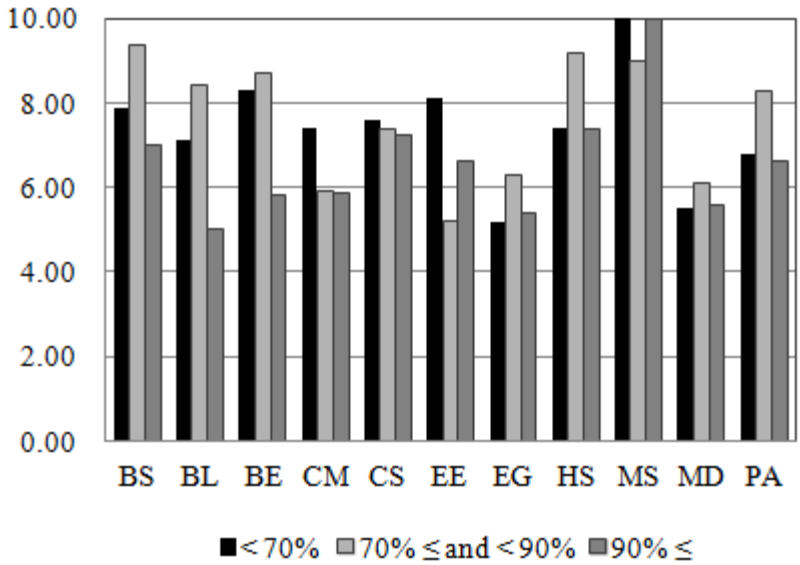


Figure 4. CHL by field and layer of cumulative ratio of access count.

Correlations between Access and Citations

In order to see the degree to which the obsolescence of citations and access are in accordance, observations were made of the correlations between (A) CHL and DHL; and between (B) II/IF and DII/DIF, respectively (Figs. 5 and 6). The distributions of II/IF and DII/DIF have high values of skewness (4.86 and 5.16, respectively). Moreover, as mentioned previously, we cannot obtain exact values for CHL from JCR, where the maximum value of CHL is 10, with values above this also being treated as 10. Thus, Spearman's rank correlation coefficient ρ is determined instead of the product–moment correlation coefficient, which should be applied to interval scales or ratio scales compliant to normal distribution. Table 2 shows the correlation coefficients for (A) CHL vs. DHL and (B) II/IF vs. DII/DIF by field.

The correlation coefficients for all 11 fields were (A) $\rho = 0.44$ and (B) $\rho = -0.02$, that is, no strong correlation was observed. In particular, almost no correlation was found with (B), which was not significant at a standard of 5%. With regard to individual fields, in the case of (A), somewhat strong and significant correlations of 0.6 or higher were seen in "Chemistry and Materials Science" ($\rho = 0.69$) and "Engineering" ($\rho = 0.68$) ($p < 0.05$). In the case of (B), the degree of correlation was weak in all fields. Furthermore, as for (B), there were not only positive correlations but also negative correlations. Among these, relatively strong and statistically significant correlations were observed in "Biomedical and Life Sciences" ($\rho = 0.35$) ($p < 0.05$).

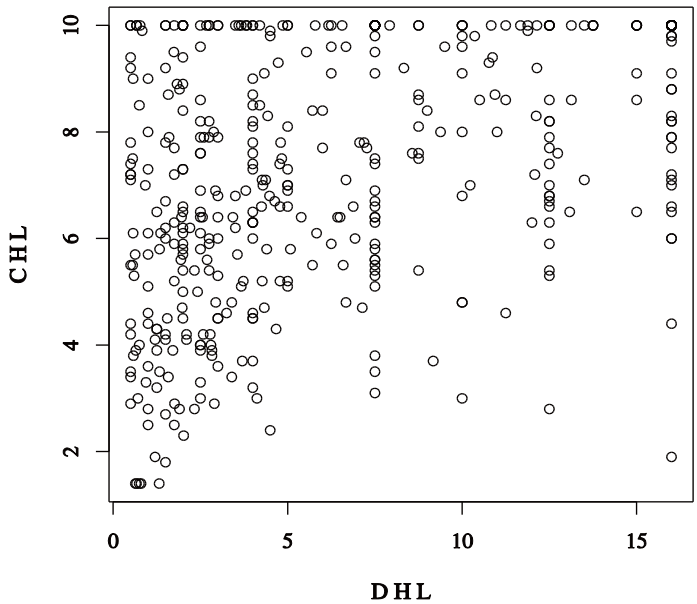


Figure 5. (A) Scatter diagram for CHL vs. DHL.

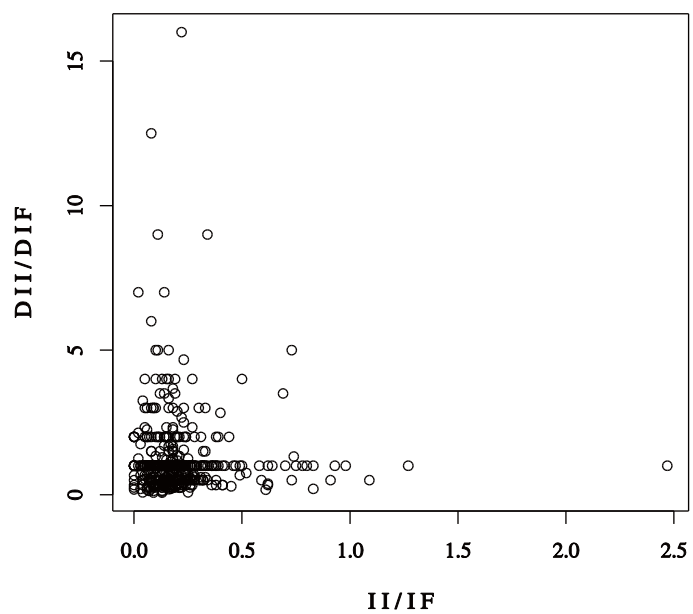


Figure 6. (B) Scatter diagram for II/IF vs. DII/DIF.

Table 2. The degree of correlation of obsolescence between citations and access.

Subject of Springer	*A*		*B*	
Behavioral Science (BS)	0.04		-0.10	
Biomedical and Life Sciences (BL)	0.31	*	**0.35**	*
Business and Economics (BE)	0.29		-0.02	
Chemistry and Materials Science (CM)	**0.69**	*	0.09	
Computer Science (CS)	0.30		-0.03	
Earth and Environmental Science (EE)	0.61	*	-0.13	
Engineering (EG)	**0.68**	*	0.07	
Humanities, Social Sciences and Law (HS)	0.06		-0.27	
Mathematics and Statistics (MS)	0.39	*	-0.21	
Medicine (MD)	0.37	*	-0.11	
Physics and Astronomy (PA)	0.47	*	0.14	
Whole	0.44	*	-0.02	

* Significant ($p < 0.05$)

Conclusions

The results of this research reveal a tendency for the obsolescence of citations and access to be comparatively long for both "Mathematics and Statistics" and "Behavioral Science." In contrast, it is short in the natural science fields other than "Mathematics and Statistics." These results are largely in line with Hypothesis 1. Note that in regard to journals belonging to the layer with the lowest access count, it was observed that obsolescence was fast in many fields.

With respect to correlations between the obsolescence of citations and access, a trend was observed in which these differ depending on the field and index. As for long term obsolescence (CHL and DHL), the degree of correlation differed by field, but all fields demonstrated a positive correlation. In other words, fields with fast obsolescence of citations witnessed a tendency toward fast obsolescence of access. Meanwhile, with respect to short term obsolescence (II/IF and DII/DIF), no consistent trend was observed. Whether the correlation was positive or negative depended on the field.

In most fields, the correlation between obsolescence of citations and that of access was stronger in the long term (CHL and DHL) than in the short term (II/IF and DII/DIF). However, in the cases of "Biomedical and Life Sciences," "Humanities, Social Sciences and Law," and "Behavioral Science," the correlation of the latter was stronger. This result was posited in Hypothesis 2, in which indices with a strong correlation differ depending on the field. Furthermore, while only weak correlations were observed in most fields for both the long term and the short term, the fields of "Engineering" and "Chemistry and Materials Science" showed somewhat strong correlations of 0.6 or higher regarding long term obsolescence. In other words, this result suggests that with respect to these two fields, it is

possible to a certain degree to predict the long term obsolescence of access on the basis of the value of CHL obtained from JCR.

Therefore, these trends provide suggestions for introducing effective journal backfiles. However, the results of this research have direct relevance only to the situation at a particular university. If we are to obtain more general insights, future research needs to incorporate surveys that are broader in scope and that organize these trends based on the size and type of university (e.g., science- versus humanities-oriented universities, or research- versus education-oriented universities).

References

Bollen, J. & van de Sompel, H. (2008). Usage impact factor: The effects of sample characteristics on usage-based impact metrics. *Journal of the American Society for Information Science and Technology*, 59(1), 136-149.

Brody, T., Harnad, S. & Carr, L. (2006). Earlier web usage statistics as predictors of later citation impact. *Journal of the American Society for Information Science and Technology*, 57(8), 1060-1072.

Chu, H. & Krichel, T. (2007). Downloads vs. citations in economics: Relationships, contributing factors & beyond. In: Proceedings of the 11th International Society for Scientometrics and Informetrics Conference, pp. 207-215, Madrid, Spain, June 25-27.

Duy, J. & Vaughan, L. (2006). Can electronic journal usage data replace citation data as a measure of journal use? An empirical examination. *The Journal of Academic Librarianship*, 32(5), 512-517.

Kurtz, M. J., Eichhorn, G., Accomazzi, A., Grant, C., Demleitner, M., Murray, S. S., Martimbeau, N. & Elwell, B. (2005). The bibliometric properties of article readership information. *Journal of the American Society for Information Science and Technology*, 56(2), 111-128.

McDonald, J. D. (2007). Understanding journal usage: A statistical analysis of citation and use. *Journal of the American Society for Information Science and Technology*, 58(1), 39-50.

Moed, H. F. (2005). Statistical relationships between downloads and citations at the level of individual documents within a single journal. *Journal of the American Society for Information Science and Technology*, 56(10), 1088-1097.

Nicholas, D., Huntington, P., Dobrowolski, T., Rowlands, I., Jamali, M.H.R. & Polydoratou, P. (2005). Revisiting 'obsolescence' and journal article 'decay' through usage data: An analysis of digital journal use by year of publication. *Information Processing and Management,* 41(6), 1441-1461.

Schloegl, C. & Gorraiz, J. (2010). Comparison of citation and usage indicators: The case of Oncology journals. *Scientometrics*, 82(3), 567-580.

Schloegl, C. & Gorraiz, J. (2011). Global usage versus global citation metrics: The case of Pharmacology journals. *Journal of the American Society for Information Science and Technology,* 62(1), 161-170.

Wan, J.-K., Hua, P.-H., Rousseau, R. & Sun, X.-K. (2010). The journal download immediacy index (DII): Experiences using a Chinese full-text database. *Scientometrics*, 82(3), 555-566.
Watson, A. B. (2009). Comparing citations and downloads for individual articles. *Journal of Vision*, 9(4), 1-4.

USING MONTE CARLO SIMULATIONS TO ASSESS THE IMPACT OF AUTHOR NAME DISAMBIGUATION QUALITY ON DIFFERENT BIBLIOMETRIC ANALYSES.

Jan Schulz[1]

[1] *jan.schulz@bwl.tu-freiberg.de (ORCID: 0000-0002-2312-0588)*
TU Bergakademie Freiberg, Lessingstr. 45, 09599 Freiberg (Germany)

Abstract
Bibliometric analyses depend on the quality of data sets and the author name disambiguation process which attributes written papers with names on it to real persons. Errors of the author name disambiguation process can distort the results of the analyses. To assess the resulting error in the analyses outcomes, Monte Carlo simulations can be used. This paper presents a basic algorithm of such simulations and how errors will lead to changes to the results of different kinds of analyses (rankings and regressions analysis with number of papers as dependent variable).
The results show that rakings of authors are more depended on data set quality than regression coefficients. Both mean and individual per person data set quality is important for valid ranking results. Regression coefficients change less than 10% under current automatic attribution processes quality.

Conference Topic
Old and New Data Sources for Scientometric Studies: Coverage, Accuracy and Reliability (Topic 2) and Scientometrics Indicators: Criticism and new developments (Topic 1).

Introduction
Bibliometric data sets are the basis for several different analyses, ranging from rankings of authors and institutions up to inferring—combined with other data sets—the effects of author or other characteristics on research productivity. As such, publication performance measurements are important to evaluate scientists and research institutions and base policy decisions on it (e.g. Abramo & D'Angelo, 2011; Frey, 2003)
All such analysis depend on the quality of the underlying data sets: errors in the data set will affect the results of these analyses and potentially wrong conclusions and decisions could be drawn from these wrong analyses results.

Problems of assessing the impact of errors in the data
Manually cleaning up the underlying bibliometric data sets is time and resource consuming and with millions of articles not feasible (Wang et al., 2012). As such it is important to know the impact of errors in bibliometric data sets on the

resulting analysis result. Such assessments of the impact can be done by using a sample of the data set and manually checking and cleaning up this sample before comparing the analyses results from both data sets. This process is itself time and resource consuming and—as a manual step is involved—potentially still not error free. This paper will introduce Monte Carlo simulations as a different mean to assess the impact of errors in bibliometric data sets on analyses results.

Data and Methods

Attribution process and error measurements

Errors in the bibliometric data set can have two sources: (1) the underlying raw data sets (e.g. *Web of Science* or *Scopus*) can contain typos, omissions, or other errors or (2) the process which attributes a paper to an entity (e.g. an author) wrongly attributes it to the wrong entity. This process is known as author name disambiguation process (ANDP). The results of both errors are that some entities have more and others have less than their real share of articles attributed to them. To mitigate such problems more signals than only the author names are used in the ANDP, such as the email address or even web searches (see Table 11 for examples), but even the most sophisticated ANDP is not error free.

The quality of an ANDP and the resulting data-set can be measured by *precision*, *recall* and *F1 scores* (Heath et al., 2009; Torvik & Smalheiser, 2009). *Precision* is the ratio between correctly attributed and all attributed papers and *recall* the ratio between correctly attributed papers and all papers written by a person. The harmonic mean of both indicators is called *F1-score*. All indicators are between zero and one with high values signaling better quality.

Table 11: Reported quality measures of different author name disambiguation processes.

F1 score (±95% CI)	*Type*	*Dataset and method details*
0.36±0.05 – 0.46±0.14	Unsupervised learning	Top 14 common names in the DDPL data set, automatic machine learning methods (clustering, SVM) (Heath et al., 2009)
0.76±0.08	Web searches as signal	Top 14 common names in the DDPL data set (Heath et al., 2009)
0.767±0.060	„Self training“	Survey of authors, auto generated training set (co-authors, email) (Levin et al., 2012)
0.953	Manual reference list	Italian National Citation Report (D'Angelo, Giuffrida, & Abramo, 2011)

Table 1 shows mean F1 scores from different ANDP implementations and their 95% confidence interval (if available or derivable), which were measured by comparing the results of the ANDP against a manually cleaned sample. Quality

varies wildly and increases with more signals and manual training steps. Quality is between 0.36 and 0.77 for fully automatic ANDP and better than F1=0.90 for ANDPs with reference lists or manual attribution steps. The confidence intervals show that these quality measurements differ greatly for individual researchers.
For commercial available data sets such as the *Web of Science* or *Scopus* no such *F1 scores* could be found but Torvik & Smalheiser (2009, p. 21) reports that for "139 cases […] Web of Science and Scopus split 7.8% (11/139) and 18.7% (26/139) of [pairs of papers from the same author] into separate clusters", meaning that both databases have high false negative counts.

Monte Carlo simulations

To assess the impact of such errors, Monte Carlo simulation can be used. Monte Carlo simulations use "random sampling and statistical modeling to estimate mathematical functions and mimic the operations of complex systems" (Harrison, 2010, p. 17). Instead of mathematically deriving the statistical distribution of the error (the differences between analysis results of the *real* and the *measured* data set) from the distribution of the errors in the data set, Monte Carlo simulations first simulate the data set and the errors and afterwards apply the analysis to both the *real* and the *measured* data set and the differences of the resulting analyses are counted. Doing this a few hundred or thousand times produces the distribution of the differences between both analysis results.

Simulated data sets

Two different bibliometric analyses are looked at: The use of author productivity in the form of *papers-per-author* values for (1) rankings of authors and (2) as an input in regression analysis. The simulations were done with a range of data quality.
As a first step, the real productivity values (written papers per author) of a list of authors were simulated (ppa_{real}). In real world data sets, ppa_{real} follows a power law distribution (Clauset, Shalizi, & Newman, 2009), meaning there a lots of authors with just a one or a few papers but only a few with hundreds of papers. For the analysis of the impact of errors on rankings, this value was generated directly from a power law distribution. For the analysis of the impact of errors on regression analysis the productivity of an author was assumed to be dependent on two different observed author characteristics (e.g. "network position", normal distributed with $\mu = 2$ and $\sigma = 2$) and one non-observed variable *noise* which subsumes all other influences (normally distributed with $\mu = 0$ and $\sigma = 0.5$). The final values were computed as $ppa_{real,reg} = e^{char_a + 2char_b + noise}$. This value is log normal distributed which has similar properties as a power law distribution and in real world data sets it is usually hard to distinguish from a power law distribution (Stringer, Sales-Pardo, & Nunes Amaral, 2008, p. 2).

In a second step, errors, based on mean F1 score ($F1_{mean}$) and the variance of individual F1 scores ($F1_{sigma}$), were introduced to these ppa_{real} values to derive measured values.

For the purpose of this study, it was assumed that *precision* and *recall* were equally valued resulting in the same mean value for each quality indicator (*F1 score = precision = recall*).

Precision and *recall* values per author were normally distributed with $\mu = F1_{mean}$ and $\sigma = F1_{sigma} \times \min(F1_{mean}, 1 - F1_{mean})$. σ was specified relative to the maximal possible change as e.g. a mean data quality of $F1 = 0.9$ of an individual data sets for one authors means that this data sets *F1 score* can only gain 0.1 before reaching perfect data quality. Some generated values were outside of the range [0..1] and had to be cut at these levels.

From these individual *precision* and *recall* values, falsely attributed (fp) and falsely not attributed (fn) counts per author were derived. The measured productivity was computed as $ppa_{measured} = ppa_{real} + fp - fn$.

Both $F1_{mean}$and $F1_{sigma}$ were varied in the simulations. $F1_{mean}$ ranged from 0.5 to 1.0, meaning the mean data quality ranged from "about half the observed papers are wrongly attribute and half of the real one are not attributed" to "perfect data quality". $F1_{sigma}$ ranged from 0.0 to 0.5. For each different quality specification 500 simulation runs were performed with 2500 persons each.

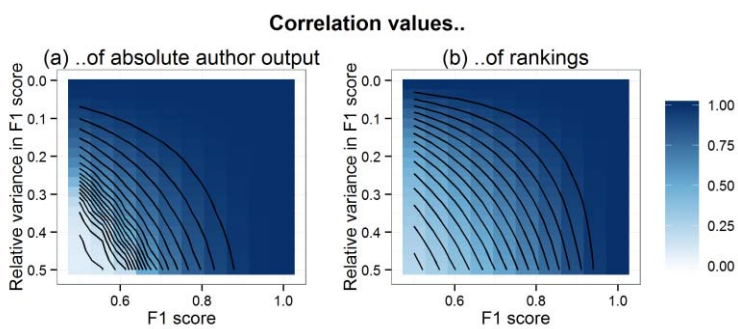


Figure 2: Correlation between (a) absolute values for author productivity and (b) between rankings of authors based on author productivity (both measured in *papers-per-author*). Each field stands for a specified data quality and is the mean correlation value of 500 simulations of 2500 authors. Contour lines represent equal correlation values and are spaced 0.05 apart.

Sensitivity analyses

To measure the sensitivity of the results of the analysis regarding different data set qualities, multiple comparison methods were employed: the overall effect was assessed by correlating the ppa_{real} values with its $ppa_{measured}$ counterparts. To assess the effect of data set quality on rankings, the correlation between both rankings was used (Wang et al., 2012, p. 400) and additionally the changes in the Top-10 list were tallied. For regression analysis $\log(ppa_{measured})$ was regressed

on the simulated author characteristics. The standard deviation as well as the range of observed regression coefficients was used as an indicator of the impact of errors on the resulting regression analysis (Cortez & Embrechts, 2013, p. 3).

Findings

Impact on author rankings

Figure 2a shows the correlation between real and measured absolute papers-per-author values for a range of data qualities (specified in mean F1 scores and variance of individual F1 scores). The upper right corner represents perfect data quality and the resulting correlation is 1.0. The correlations become lower for both lower F1 scores and higher variance in individual F1 scores.

Figure 2b shows the correlation between rankings for different data qualities. Compared to the correlation values of absolute values the correlation value for rankings is lower for higher data qualities but does not drop down as fast as the correlation values of absolute values. Both, mean data quality and individual person's data quality are important: both contribute roughly equally to the reduced correlations between real and measured papers-per-author values.

The difference between absolute and rank correlation is shown in Figure 3 for relative variance of 0.19, which corresponds roughly to the best automatic ANDPs as shown in Table 1. Correlations of rankings are consistently lower for the whole range of F1 scores.

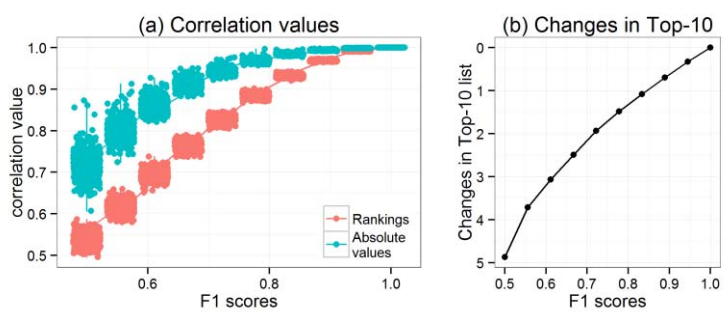


Figure 3: Detailed view at relative variance of 0.19 and with varied F1 scores: (a) correlation values for rankings and absolute papers-per-author values; (b) changes in the Top-10 ranking.

To get a better picture what these correlation values mean for rankings, the mean changes to the Top-10 list for different mean F1 scores at relative variance of 0.19 are shown in Figure 3(b). At a low data quality of $F1 = 0.5$, almost half of the persons in the Top-10 list are changed. Even at a data quality of $F1 = 0.9$, on average 0.5 persons are changed. Using 0.76 as the mean quality of a data set of the best fully automatic ANDP, on average 1.5 persons are changed in the Top-10 list.

Impact on regression analysis

Figure 4 shows the results of the analysis of the impact of data set quality on the results of a regression analysis. The displayed regression coefficient had a simulated real value of "1". Figure 4a shows the standard deviation of this regression coefficient at different data qualities. With a maximal standard deviation of 0.01 the resulting regression coefficients deviate little from the real value on the simulated range of data qualities. Looking at the maximal difference between individual regression coefficients (Figure 4b), the maximal difference is 0.07, meaning that even in the worst case, the regression coefficients are only off by 7%.

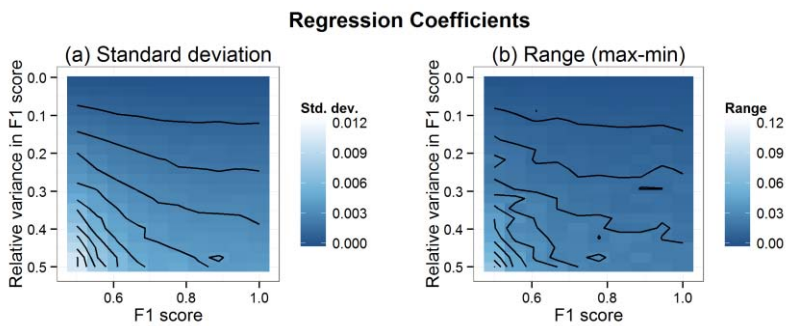


Figure 4: Properties of regression coefficients with a simulated real value of "1" at different data qualities: (a) Standard deviation of regression coefficients; (b) maximal difference of regression coefficients (max. value – min. value). Each field stands for a specified data quality and is based of 500 simulations of 2500 authors.

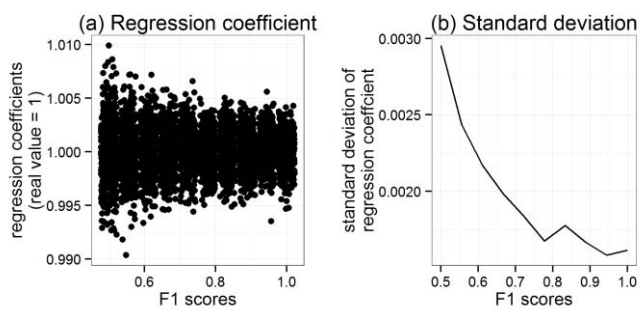


Figure 5: Detailed view at relative variance of 0.19 and at different F1 scores: (a) regression coefficient (real value = 1); (b) standard deviation of regression coefficient.

In contrast to the results of the analysis of rankings, at least for better data set quality, regression coefficients are more influenced by relative variance of individual data quality measures.

Figure 5 shows again a detailed view of different mean F1 scores at relative variance of 0.19. The maximal range between individual regression coefficients (Figure 5a) is 0.02 and declines with better F1 scores. The simulated regression coefficients are off by at most 1%. Looking at the standard deviation of regression coefficients, deviation is lower with better mean data set quality (Figure 5b) and drops from 0.0030 to about 0.0016.
Looking at F1 score = 0.76 (the best fully automatic ANDP) the simulated regression coefficients are off by at most 0.5% and have a standard deviation of less than 0.2%.

Discussion and Conclusion

The aim of this study was to show the application of Monte Carlo simulation to assess the impact of errors in bibliometric data sets on analysis results.
The analysis of the impact of data quality on rankings shows, that data quality is an important factor for the validity of such rankings. For data set qualities comparable with data set quality produced by the best fully automatic ANDP, about 1.5 persons are changed in the Top-10 list of such rankings.
It was also shown that both mean data quality and variance in individual data quality are important. As such it is important that ANDP implementations and analyses based on such data sets report both the mean data quality measures (*precision*, *recall*, and/or *F1 scores*) and the *variance of individual quality measures* for each person.
Possible problems arise from mixing authors from different cultures/ countries, as some countries have a highly skewed distribution of names (e.g. Korean names; Aksnes, 2008) or names with a higher probability of typos (e.g. German umlauts; Fenner, 2011). This implies different types of errors (lower *precision* or lower *recall* measures, respectively) for authors from these countries. Both effects lower the mean data quality as well as imply differences in individual data quality, which would reduce the validity of rankings.
In the analysis of the impact of data quality in regression analysis, the underlying data set quality has almost no impact on the resulting regression coefficient, even for bad data qualities. For data set qualities of the best fully automatic ANDP, the observed regression coefficients were off by at most 0.5%.
As the simulated errors in the data set are not correlated to the independent variables (characteristics of authors) they only add more noise to the regression model but do not influence the value of the regression coefficients. The implication is that regression analysis with bibliometric data sets should test for systematic differences in errors which are correlated to independent variables (as would be the case of country as an independent variable) and which could influence the validity of the regression results.
The adaption a unique identifier for scholarly authors, such as of ORCID (http://about.orcid.org/), will make ANDP obsolete and therefore the problem of bibliometric data set quality for rankings will hopefully be reduced (Fenner, 2011).

References

Abramo, G., & D'Angelo, C. A. (2011). Evaluating research: from informed peer review to bibliometrics. *Scientometrics*, *87*(3), 499–514. doi:10.1007/s11192-011-0352-7

Aksnes, D. W. (2008). When different persons have an identical author name. How frequent are homonyms? *Journal of the American Society for Information Science and Technology*, *59*(5), 838–841. doi:10.1002/asi.20788

Clauset, A., Shalizi, C. R., & Newman, M. E. J. (2009). Power-Law Distributions in Empirical Data. *SIAM Review*, *51*(4), 661. doi:10.1137/070710111

Cortez, P., & Embrechts, M. J. (2013). Using sensitivity analysis and visualization techniques to open black box data mining models. *Information Sciences*, *225*(0), 1–17. doi:10.1016/j.ins.2012.10.039

D'Angelo, C. A., Giuffrida, C., & Abramo, G. (2011). A heuristic approach to author name disambiguation in bibliometrics databases for large-scale research assessments. *Journal of the American Society for Information Science and Technology*, *62*(2), 257–269. doi:10.1002/asi.21460

Fenner, M. (2011). Author Identifier Overview. *LIBREAS.Library Ideas*, *7*(1), 24–29.

Frey, B. S. (2003). Publishing as prostitution? - Choosing between one's own ideas and academic success. *Public Choice*, *116*(1/2), 205–223. doi:10.1023/A:1024208701874

Harrison, R. L. (2010). Introduction to Monte Carlo Simulation. *AIP Conference Proceedings*, *1204*, 17–21. doi:10.1063/1.3295638

Heath, F., Rice-Lively, M. L., Furuta, R., Pereira, D. A., Ribeiro-Neto, B., Ziviani, N., … (2009). Using web information for author name disambiguation. In *Proceedings of the 2009 joint international conference on Digital libraries - JCDL '09* (pp. 49–58). ACM Press.

Levin, M., Krawczyk, S., Bethard, S., & Jurafsky, D. (2012). Citation-based bootstrapping for large-scale author disambiguation. *Journal of the American Society for Information Science and Technology*, *63*(5), 1030–1047. doi:10.1002/asi.22621

Stringer, M. J., Sales-Pardo, M., & Nunes Amaral, L. A. (2008). Effectiveness of Journal Ranking Schemes as a Tool for Locating Information. *PLoS ONE*, *3*(2), e1683 EP -. doi:10.1371/journal.pone.0001683

Torvik, V. I., & Smalheiser, N. R. (2009). Author name disambiguation in MEDLINE. *ACM Transactions on Knowledge Discovery from Data*, *3*(3), 1–29. doi:10.1145/1552303.1552304

Wang, D. J., Shi, X., McFarland, D. A., & Leskovec, J. (2012). Measurement error in network data: A re-classification. *Social Networks*. doi:10.1016/j.socnet.2012.01.003

VISUALIZING AND COMPARING THE DEVELOPMENT OF SCIENTIFIC INSTRUMENTATION VS ENGINEERING INSTRUMENTATION

Chunjuan Luan[1], Xianwen Wang[2] and Haiyan Hou[3]

[1] *julielcj@163.com*
Dalian University of Technology, School of Administration & Law, WISE Lab, Dalian City Linggong Road 2#, 116023 Dalian (P. R. China)

[2] *xianwenwang@dlut.edu.cn*
Dalian University of Technology, School of Administration & Law, WISE Lab, Dalian City Linggong Road 2#, 116023 Dalian (P. R. China)

[3] *htieshan@dlut.edu.cn*
Dalian University of Technology, School of Administration & Law, WISE Lab, Dalian City Linggong Road 2#, 116023 Dalian (P. R. China)

Abstract
Laboratory instrumentations are considered a significant factor leading to innovation. But what are the developing trends like of Scientific Instrumentation (SI) and Engineering Instrumentation (EI)? Are their trends also appear to be exponential law? How many stages are existing in SI & EI developing process? what are the network structures like of SI & EI? What are the similarities and differences between SI & EI in terms of their trends, stages, and network structures? Few related studies have been found yet, and this study intends to explore the above questions by using patent analysis and social network analysis, hoping get an overview concerning the characteristics of SI & EI developing during 1963-2011. Results show that both developing trend curves of SI & EI are striking similar, tend to be polynomial developing trend, and have striking similar stages clustered and separated by SPSS. Comparing network structures of SI & EI (2011) shows that, Electrical and Electronic is the biggest sub-network for both networks. Comparing top subject areas of SI & EI shows that, Chemistry, is the biggest Subject Area in SI, and Computer Science is the first Subject Area in EI. Still some related questions should be further explored.

Conference Topic
Technology and Innovation Including Patent Analysis (Topic 5) and Visualisation and Science Mapping: Tools, Methods and Applications (Topic 8).

Introduction
It is widely acknowledged that SI & EI are vital to science & technology (S&T) development. But what are the developing trends like of SI & EI? Are their trends

also appear to be exponential law? How many stages are existing in SI & EI developing process? what are the network structures like of SI & EI? What are the similarities and differences between SI & EI in terms of their trends, stages, and network structures? Few related studies have been found yet, and this study intend to explore the above questions, hoping get an overview concerning the characteristics of SI & EI developing.

In this study, Scientific Instrumentation (SI) and Engineering Instrumentation (EI) are included in Laboratory Equipment (LE). SI is mainly used in scientific research, and EI is chiefly applied in engineering practice. SI & EI are two parallel sub-fields belonging to the upper field of Instrumentation-Measuring-Testing in the so called World Patent Database of Derwent Innovations Index (DII).

The existing studies are mainly pertinent to SI, few concerning EI. Studies related to SI include the following facets. Firstly, determinants to SI innovation. Hippel (Hippel 1976) had a sample of one hundred and eleven scientific instrument innovations studied to determine the roles of instrument users and instrument manufacturers in the innovation processes which culminated in the successful commercialization of those instruments, and found that approximately 80% of the innovations judged by users to offer them a significant increment in functional utility were in fact invented, prototyped and first field-tested by users of the instrument rather than by an instrument manufacturer. Bergen and Pearson (Bergen and Pearson 1983) argued project management and innovation in the scientific instrument industry.

Secondly, innovations between SI users and SI manufacturers. Riggs and Vonhippel (Riggs and Vonhippel 1994) also explore the relationship between the sources of innovation and incentives to innovate in a sample of 64 innovations related to Auger and Esca, two types of scientific instrument used to analyze the surface chemistry of solid materials. And found that innovations with high scientific importance tend to be developed by instrument users, while innovations having high commercial importance tend to be developed by instrument manufacturers.

Some related studies concerning SI have also been explored, such as analyzing practices and achievements in United States-Kingdom and West Germany in the scientific instrument industry (Bergen 1982); scientific instruments as keys to artificial revelation (Beaver 1988), and the relationship of scientific instruments, scientific progress and the cyclotron (Baird and Faust 1990), et al..

The extant studies are mainly conducted by using a considerably smaller sample, and discuss some issues related to determinants to SI innovation or difference innovation between SI users and manufacturers, et al.. No related studies concerning the developing trends, stages and network structures of SI & EI have been found, yet.

Patent data from DII are selected to visualize and compare the developing trends, stages and network structures of SI & EI in this study. For patent documents contain rich technical information related to intellectual property rights and

important research results (Lawson, Kemp et al. 1996; Tseng, Lin et al. 2007; Magerman, Van Looy et al. 2010), and are usually considered the proper datasets in the analysis of technology and industry developing.
This paper is organized as follows: Following this introduction, Section 2 introduces related studies. Section 3 presents the dataset and methodologies in this study. Section 4 visualize and compare trends, stages and network structure of SI & EI. Section 5 concludes this study and discusses the findings and the implications.

Literature review

On developing trends of SI & EI

After Price's exponential growth constant was presented, the exponential growth law has been tested widely, especially in science development and related domains. Leydesdorff and Zhou (Leydesdorff and Zhou 2005) proposed that in China and Korea, in addition to publications, their citation rates keep pace with the exponential growth patterns, albeit with a delay. Furthermore, some related studies have revised Price's exponential law, for example, Su and Han (Su and Han 1998) replaced by a polynomial of degree n-1 to Price's exponential law, and their research showed that the new model was more convincing than the former ones, and they also gave detailed calculation procedure, examples, parameter values and mean square errors. What are the developing trends like of SI & EI? Are their trends also appear to be exponential law, or others? No related studies have been found yet.

On developing stages of SI & EI

Phasing or stage division is usually the fundamental work in bibliometric researching. Averaging method (Hou, Kretschmer et al. 2008) or visual method (Makovetskaya and Bernadsky 1994) is generally employed in stage dividing. Studies concerning stages division of SI & EI have not been found. Therefore, a new method by using SPSS (Statistical Package for the Social Sciences) will be tried to cluster and separate developing stages of SI & EI.

On network structures of SI & EI

Social network analysis (SNA) is not a formal theory in sociology but rather a strategy for investigating social structures (Otte and Rousseau 2002). SNA has widely employed in scientometrics, such as collaboration analysis (Kretschmer and Aguillo 2004; Wagner and Leydesdorff 2005; Schilling and Phelps 2007; Hou, Kretschmer et al. 2008; Leydesdorff and Wagner 2008); co-citation analysis (Wouters and Leydesdorff 1994; Otte and Rousseau 2002; Marion, Garfield et al. 2003; Johnson and Oppenheim 2007; Chen, Chen et al. 2009; Leydesdorff 2009; Wang, Zhang et al. 2011); co-occurrence analysis (Small 1973; Morris 2001; Leydesdorff and Vaughan 2006; Leydesdorff 2007; Waltman and van Eck 2007; Jeong and Kim 2010), et al.. However, only a few studies on technology network

by using SNA have been found. The existing studies focus on the influence of business strategies on technological network activities (Gemunden and Heydebreck 1995; Vanhaverbeke, Gilsing et al. 2012); high-technology network in northern Finland (Jauhiainen 2006); the role of transnational corporations in the Chinese science and technology network (Hennemann 2011); global technology analysis (Nam and Barnett 2011), et al.. Studies concerning network structure of SI & EI by using patent analysis have not been found.

Data and Methodology

Data in the study

1. Data retrieval

The data in this study is retrieved from Derwent Innovations Index (abbr. as DII below). DII is one of the most comprehensive databases collecting patent documents in the world, begun in 1963 and currently published by the Thomson Reuters. DII includes three parts: Chemistry, Engineering and Electric & Electronics. Every week 25, 000 patent documents published by more than 40 patent offices and 45, 000 patent citation documents from 6 important copyright offices are input into DII. The date used in our statistics has been officially published. Because one patent family include one basic patent and one or more equivalent patent(s), the number of patents in this study is of the basic patent, not all applications. The basic patent publication date is definite.
Each bibliographic patent record in DII is assigned with three different classification standards: International Patent Classification (IPC), Derwent Class Code (DC), and Derwent Manual Code (MC). DC is used to retrieve the data of SI & EI, for among DCs, S02 represents Engineering Instrumentation and S03 represents Scientific Instrumentation. Time span=1963-2011; database=CDerwent, EDerwent, MDerwent.

2. Data format converting and processing

Data downloaded from DII should be converted first. Data format conversion function of CiteSpace (Chen 2006) is employed to convert the patent publications into the web of science export format, for most of the data-processing research software such as CiteSpace, Bibexcel, et al. being designed originally to process data with the format of Web of Science, and now CiteSpace also provides data format conversion function for several other kinds of data downloaded from PubMed, arXiv, ADS, and NSF Award Abstracts et al. (Chen 2010). Bibexcel (Persson 2012) and Ucinet (Jang 2000) are applied to analyze the converted patent data of SI & EI.

Separating developing stages of SI & EI

1. Selection of variables

Separating developing stages is usually the fundamental and premise work before doing some research. Selecting variables is the necessary preparation when employing SPSS to cluster and separate developing stages. Three variables are always the minimum requirement. These variables should operate a same standard, and change continuously. For example, when SI developing stages is separated, different years and different number of patent filings in different years are two proper variable, but how to choose the third variable is not easy. Different number of DC (Derwent Class Code) in different years is selected as the third variable at last, after comparing with other variables. The main reason of choosing DC is DC operates a unified standard other than IPC, MC, et al. which operate a hierarchical classification system.

2. Hierachical Cluster Analysis

According to at least three different variables, employing Hierachical Cluster Analysis embedded in the software of SPSS (Statistical Package for the Social Sciences), selecting cluster method of Between-groups linkage, and measured by Squared Euclidean Distance, choosing a proper rang of solutions, then several possible solutions can be got, and a best result could be picked out.

Visualizing network structure

According to the methodologies and procedures elaborated in the article of "Mapping the evolution of technology network in the field of solar energy technology" (Luan, Hou et al. 2012) , networks of SI & EI are drawn out respectively, and their network structures are compared with each other.

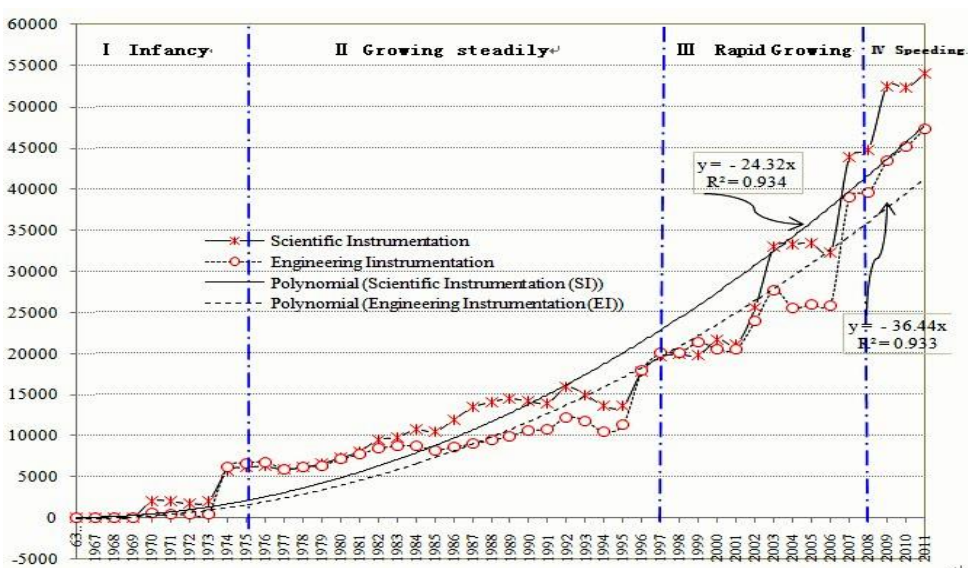


Figure 1. Developing trend curves and stages of SI & EI during 1963-2011.

Analysis and results

Comparing the developing trends of SI & EI

During 1963-2011, total patent filings of SI are 766, 122, and those of EI are 657, 365, existing a gap of 108, 757. The developing trends of SI & EI are shown in Figure 1.

The two developing trend curves are striking similar in Figure 1. Both of them demonstrate an increasing developing trend obviously over time, with fluctuating a bit during their going up ways. Differing from the exponential growth law of scientific development, the two curves of SI & EI are tend to be polynomial developing trend, with goodness of fit of 0.934 and 0.933, respectively.

Comparing the developing stages of SI & EI

According to three different variables, that is, different publication years, different number of publications (patent filings in this paper) in different years, and different number of Derwent Class Classifications (DC), by using the Analysis Functions called Hierachical Cluster Analysis embedded in the software of SPSS (Statistical Package for the Social Sciences), all the years during 1963-2011 are clustered and separated into 4 stages. Stage Ⅰ: 1963-1973, a infancy stage, developing speed in this stage is slow; Stage Ⅱ: 1974-1995, a steady growing stage, developing speed in this stage is steady; Stage Ⅲ: 1996-2006, a rapid growing stage, in this stage, patent filings develop at a comparatively rapid speed; Stage Ⅳ: 2007-2011, a highly speeding stage, this stage enjoying an extraordinary fast speed.

Developing stages of SI & EI are clustered by using SPSS, respectively, but it is striking that the two developing stage curves are nearly the same, so they are combined together in Figure 1.

Comparing network structures of SI & EI in 2011

1. Network structure of SI in 2011

The data in the latest year of 2011 are selected to be drawn network, and network structures are compared between SI & EI.

All the patent filings of 54, 098 in 2011 are analyzed by using the hypertext software of Bibexcel. These patent filings are pertinent to 278 different technology classifications in terms of DCs. The total frequency of 278 DCs among 54, 098 records is 160, 306, mean value of DCs in each record is 2.96. 278 DCs are all chosen in doing technology co-classification analysis. After getting the co-classification matrix of the 278 technology classifications, Jaccard index is used to obtain the normalized matrix. Netdraw tool of Ucinet is employed to draw the network of SI in 2011, respectively. By adjusting the threshold continuously, we get the clear main component network (Figure 2). The size of the nodes represents the value of degree centrality in Figure 2.

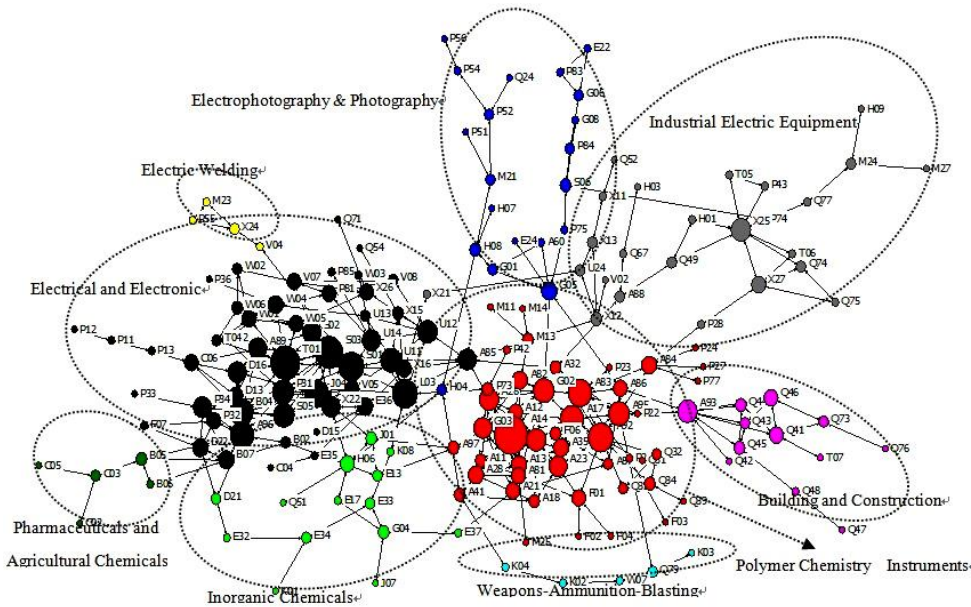


Figure 2. Network structure of Scientific Instrumentation (SI): 2011, degree.

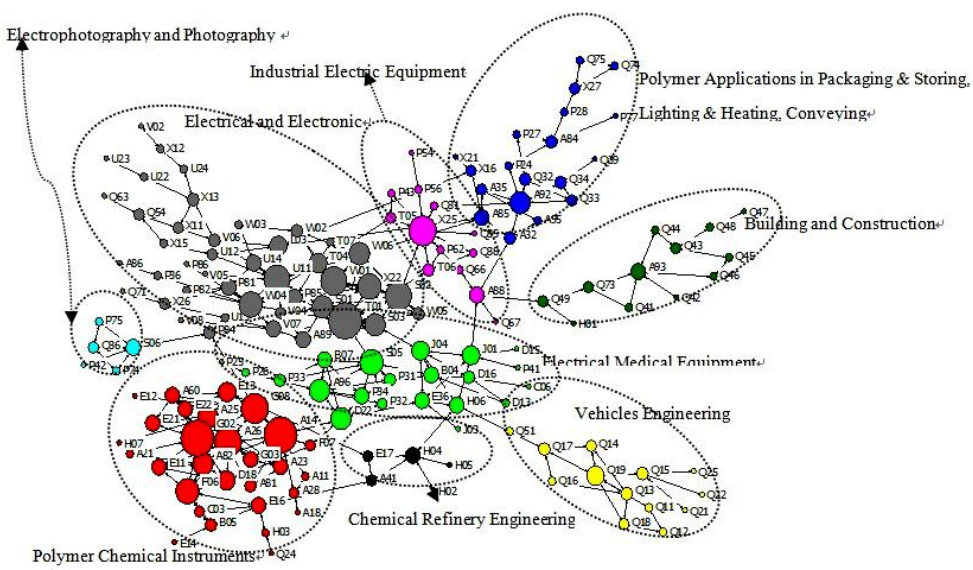


Figure 3. Network structure of Engineering Instrumentation (EI): 2011, degree.

Girvan–Newman algorithm is used to cluster sub-networks, and so the network structure of SI is demonstrated in Figure 2. *Electrical and Electronic* is the biggest sub-network, which is surrounded by several chemicals related sub-networks. The second biggest sub-network is *Polymer Chemistry Instruments*,

locating in the center of the whole network. In addition, there are several comparatively smaller sub-networks on the edge, such as *Industrial Electric Equipment, Electrophotography & Photography* and *Building and Construction*, et al..

2. Network structure of EI in 2011

By using the same steps and methods, network structure of EI in 2011 is obtained (Figure 3). The biggest sub-network is Electrical and Electronic; which is followed by another comparatively bigger sub-network named *Polymer Chemical Instruments*; and sub-networks such as *Electrical Medical Equipment, Polymer Applications, Industrial Electric Equipment, Vehicles Engineering,* et al.. And sub-networks as *Chemical Refinery Engineering and Electrophotography* and *Photography* are smaller ones.

Comparing top subject areas of SI & EI: 2011

According to the Subject Areas provided in DII, we compare top Subject Areas of SI & EI in 2011. Subject Areas with proportion ≥1% are listed in Table 1 and Table 2.

Table 1. Top Subject Areas with proportion ≥1% of SI in 2011.

Ranking	*SI: Subject Areas (Total 89)*	*% of SI*
*	*Engineering*	*99.88%*
*	*Instruments & Instrumentation*	*99.88%*
1	Chemistry	41.29%
2	Computer Science	21.27%
3	Pharmacology & Pharmacy	20.01%
4	Biotechnology & Applied Microbiology	14.22%
5	Polymer Science	12.64%
6	General & Internal Medicine	8.94%
7	Transportation	4.18%
8	Energy & Fuels	3.96%
9	Communication	2.59%
10	Food Science & Technology	2.39%
11	Agriculture	2.31%
12	Optics	1.90%
13	Imaging Science & Photographic Technology	1.76%
14	Metallurgy & Metallurgical Engineering	1.64%
15	Water Resources	1.40%

Table 2. Top Subject Areas with proportion ≥1% of EI in 2011.

Ranking	*EI: Subject Areas (Total 63)*	*% of EI*
*	*Engineering*	*99.92%*
*	*Instruments & Instrumentation*	*99.92%*
1	Computer Science	25.75%
2	Transportation	16.79%
3	Chemistry	13.05%
4	Polymer Science	7.00%
5	Communication	4.25%
6	General & Internal Medicine	4.02%
7	Energy & Fuels	3.64%
8	Metallurgy & Metallurgical Engineering	2.33%
9	Construction & Building Technology	1.98%
10	Optics	1.93%
11	Imaging Science & Photographic Technology	1.43%
12	Mining & Mineral Processing	1.27%
13	Pharmacology & Pharmacy	1.11%

It should be noted that Subject Areas of Engineering, and Instruments & Instrumentation, are excluded in analyzing, due to almost all the data are concerning these two Subject Areas. It also should be noted that total percentage exceeds 100%, because of multidisciplinary distributing of patent filings in terms of Subject Areas.

Total 89 Subject Areas are related to SI. *Chemistry*, is the biggest Subject Area in SI, with 41.29% of total in 2011; *Computer Science* is the second biggest Subject Area, with 21.27% of total; and the third one is *Pharmacology & Pharmacy*, with the proportion of 20.1%. Followed by *Biotechnology & Applied Microbiology*, and *Polymer Science*. The ratio of the top 5 Subject Areas exceeds 10%.

Total 63 Subject Areas are pertinent to EI. Differing from SI, Computer Science, is the first Subject Area in EI, with 25.75% of total in 2011; Transportation, is the second biggest Subject Area, with 16.79% of total; and the third one is Chemistry, with the proportion of 13.05%. Followed by *Polymer Science*, *Communication* and *General & Internal Medicine*. Only the top 3 Subject Areas exceeds 10% concerning proportion.

Comparing overlapping of SI & EI

Some laboratory instrumentations are used in scientific research, so they are assigned in the field of SI in DII, and at the same time, they maybe also be used in engineering, therefore they are also assigned in the field of Engineering Instrumentation in DII. This results in overlapping of SI & EI. Analyzing and

comparing the overlapping of SI & EI, will help us understand which one depends the other more during their developing process.

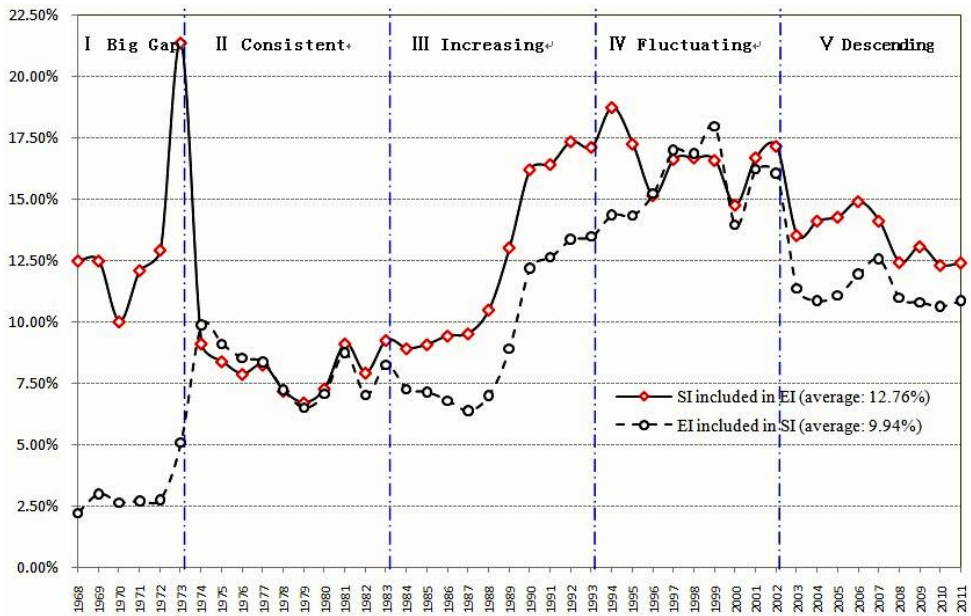


Figure 4. Overlapping of SI & EI: 1968-2011.

It should be noted that (Figure 4) no overlapping of SI & EI emerged when patent filings were less before the year of 1967; SI & EI have been overlapping since the year of 1968, so SI & EI overlapping analysis is conducted during 1968-2011.

The same, according to three different variables, that is, different years, different percentage of SI included in EI in different years, and different percentage of EI included in SI in different years, also by using the Analysis Functions called Hierarchical Cluster Analysis embedded in the software of SPSS, all the years during 1968-2011 are clustered and separated into 5 stages. Stage Ⅰ: 1968-1973, a big gap stage, the proportion of "SI included in EI" is much higher than that of "EI included in SI" in this stage, which indicating that EI depends on SI more in this period; Stage Ⅱ: 1974-1983, a comparatively consistent stage, two curves matches so well in this stage; Stage Ⅲ: 1984-1993, an increasing stage, both curves appear to be going up dominant trend in this period, and EI ; Stage Ⅳ: 1994-2002, a fluctuating stage, in this stage both curves wave fiercely; Stage Ⅴ: 2003-2011, a descending stage, two curves demonstrate decreasing in fluctuating.

Conclusions and discussions

Conclusions

Using patent data in the field of Scientific Instrumentation (SI) and Engineering Instrumentation (EI) worldwide downloaded from one of the most comprehensive world patent databases Derwent Innovation Index (DII) during 1963-2011, we visualize and compare the developing trends, developing stages, network structure, top subject areas, and overlapping of SI & EI during 1963-2011, and have an overview concerning SI & EI developing.

The two developing trend curves are striking similar, and are tend to be polynomial developing trend, with perfect goodness of fit of 0.934 and 0.933, respectively.

According to three different variables, by using the Analysis Functions called Hierarchical Cluster Analysis embedded in the software of SPSS (Statistical Package for the Social Sciences), we get another striking finding, that is, all the years during 1963-2011 are clustered and separated into 4 stages, and the developing stages of SI & EI are exactly the same, both are as follow: 1963-1973, a infancy stage; 1974-1995, a steady growing stage; 1996-2006, a rapid growing stage and 2007-2011, a highly speeding stage.

Comparing network structures of SI & EI (2011) shows the similarity is that, Electrical and Electronic is the biggest sub-network for both networks. The difference is that, as far as SI network structure, there are several chemicals related sub-networks, and some other sub-networks such as Industrial Electric Equipment, Electrophotography & Photography and Building and Construction, et al..; when it comes to EI network structure, except for the biggest sub-network, followed by another comparatively bigger sub-network named Polymer Chemical Instruments; and sub-networks such as Electrical Medical Equipment, Polymer Applications, Industrial Electric Equipment, Vehicles Engineering, et al...

Comparing top subject areas of SI & EI shows that, Total 89 Subject Areas are related to SI; total 63 Subject Areas are pertinent to EI. *Chemistry*, is the biggest Subject Area in SI, with 41.29% of total in 2011; *Computer Science* is the second biggest Subject Area, with 21.27% of total. Differing from SI, *Computer Science*, is the first Subject Area in EI, with 25.75% of total in 2011; *Transportation*, is the second biggest Subject Area, with 16.79% of total.

Comparing overlapping of SI & EI shows that, all the years during 1968-2011 concerning SI & EI overlapping are clustered and separated into 5 stages: 1968-1973, a big gap stage; 1974-1983, a comparatively consistent stage; 1984-1993, an increasing stage; 1994-2002, a fluctuating stage; and 2003-2011, a descending stage.

Discussions

Differing from the exponential law in science developing, both trends of SI & EI curves show a polynomial developing trend with fluctuations. What are the

reasons, and what impacts on science and technology developing, should be further studied.

What affects developing stages of SI & EI? And what characteristics concerning science and technology developing are resulted by the different stages? Related studies should be explored next.

Only network structures of SI & EI in 2011 have been visualized and compared. How are both network structures evolving? And what are the characteristics in different stages like? Such questions still need sufficient time to investigate.

On subject areas and overlapping of SI & EI, also need time and energy to explore in details.

Findings in this study will benefit us comprehending the regulation of SI & EI, the relationship of SI & EI, and the impacts of SI & EI on science and technology developing.

Acknowledgments

This work was supported by the Natural Science Foundation of China (NSFC) under Grant 71073015, 71103022.

References

Baird, D. and T. Faust (1990). cientific instruments, scientific progress and the cyclotron. *British Journal for the Philosophy of Science,* 41(2): 147-175.

Beaver, D. D. (1988). The secrets of science unlocked - scientific instruments as keys to artificial revelation. *Science Technology & Human Values,* 13(3-4): 373-384.

Bergen, S. A. (1982). The r-and-d production interface - united-kingdom and west-german practices and achievements in the scientific instrument industry. *R & D Management,* 12(1): 21-25.

Bergen, S. A. and A. W. Pearson (1983). P*roject-management and innovation in the scientific instrument industry*. *Ieee Transactions on Engineering Management*, 30(4): 194-199.

Chen, C. (2010). CiteSpace: 2003-2012. Retrieved October 6, 2012 from: http://cluster.cis.drexel.edu /~cchen /citespace/.

Chen, C. M. (2006). CiteSpace II: Detecting and visualizing emerging trends and transient patterns in scientific literature. *Journal of the American Society for Information Science and Technology,* 57(3): 359-377.

Chen, C. M., Y. Chen, et al. (2009). Towards an explanatory and computational theory of scientific discovery. *Journal of Informetrics*, 3(3): 191-209.

Gemunden, H. G. and P. Heydebreck (1995). The influence of business strategies on technological network activities. *Research Policy* ,24(6): 831-849.

Hennemann, S. (2011). The role of transnational corporations in the chinese science and technology network. *Erdkunde*, 65(1): 71-83.

Hippel, E. V. (1976). Dominant role of users in scientific instrument innovation process. *Research Policy*, 5(3): 212-239.

Hou, H., H. Kretschmer, et al. (2008). The structure of scientific collaboration networks in Scientometrics. *Scientometrics,* 75(2): 189-202.
Jang, Y. S. (2000). The worldwide founding of ministries of science and technology, 1950-1990. *Sociological Perspectives,* 43(2): 247-270.
Jauhiainen, J. S. (2006). Multipolis: High-technology network in northern Finland. *European Planning Studies,* 14(10): 1407-1428.
Jeong, S. and H. G. Kim (2010). Intellectual structure of biomedical informatics reflected in scholarly events. *Scientometrics*, 85(2): 541-551.
Johnson, B. and C. Oppenheim (2007). How socially connected are citers to those that they cite? *Journal of Documentation*, 63(5): 609-637.
Kretschmer, H. and I. F. Aguillo (2004). Visibility of collaboration on the Web. *Scientometrics*, 61(3): 405-426.
Lawson, M., N. Kemp, et al. (1996). Automatic extraction of citations from the text of English-language patents - An example of template mining. *Journal of Information Science*, 22(6): 423-436.
Leydesdorff, L. (2007). "Should co-occurrence data be normalized? A rejoinder." Journal of the American Society for Information Science and Technology 58(14): 2411-2413.
Leydesdorff, L. (2008). On the normalization and visualization of author co-citation data: Salton's cosine versus the Jaccard index. *Journal of the American Society for Information Science and Technology*, 59(1): 77-85.
Leydesdorff, L. (2009). How are New Citation-Based Journal Indicators Adding to the Bibliometric Toolbox? *Journal of the American Society for Information Science and Technology*, 60(7): 1327-1336.
Leydesdorff, L. and L. Vaughan (2006). Co-occurrence matrices and their applications in information science: Extending ACA to the Web environment. *Journal of the American Society for Information Science and Technology*, 57(12): 1616-1628.
Leydesdorff, L. and C. S. Wagner (2008). International collaboration in science and the formation of a core group. *Journal of Informetrics*, 2(4): 317-325.
Leydesdorff, L. and P. Zhou (2005). Are the contributions of China and Korea upsetting the world system of science? *Scientometrics,* 63(3): 617-630.
Luan, C. J., H. Y. Hou, et al. (2012). Mapping the evolution of technology network in the field of solar energy technology. 17th International Conference on Science and Technology Indicators (STI) , Montreal, Quebec, Canada. *Proceedings of STI 2012 Montreal*, 561-568.
Magerman, T., B. Van Looy, et al. (2010). Exploring the feasibility and accuracy of Latent Semantic Analysis based text mining techniques to detect similarity between patent documents and scientific publications. *Scientometrics*, 82(2): 289-306.
Makovetskaya, O. and V. Bernadsky (1994). Scientometric indicators for identification of technology system life-cycle phase. *Scientometrics*, 30(1): 105-116.

Marion, L. S., E. Garfield, et al. (2003). Social network analysis and citation network analysis: Complementary approaches to the study of scientific communication (SIG MET). Humanizing Information Technology: From Ideas to Bits and Back. Medford, Information Today Inc. Eric Archambault (Ed.), *Proceedings of the 66th Asist Annual Meeting*, 40: 486-487.

Morris, T. (2001). Visualizing the structure of medical informatics using term co-occurrence analysis: II. INSPEC perspective. *Proceedings of the 64th Asist Annual Meeting,* Medford, Information Today Inc. 38: 489-497.

Nam, Y. and G. A. Barnett (2011). Globalization of technology: Network analysis of global patents and trademarks. *Technological Forecasting and Social Change*, 78(8): 1471-1485.

Otte, E. and R. Rousseau (2002). Social network analysis: a powerful strategy, also for the information sciences. *Journal of Information Science*, 28(6): 441-453.

Persson, O. (2012). BibExcel. Retrieved October 6, 2012 from: http://www8.umu.se/inforsk/Bibexcel/.

Riggs, W. and E. Vonhippel (1994). Incentives to innovate and the sources of innovation - the case of scientific instruments. *Research Policy*, 23(4): 459-469.

Schilling, M. A. and C. C. Phelps (2007). Interfirm collaboration networks: The impact of large-scale network structure on firm innovation. *Management Science*, 53(7): 1113-1126.

Small, H. G. (1973). Relationship between citation indexing and word indexing - study of co-occurrences of title words and cited references. *Proceedings of the American Society for Information Science*, 10: 217-218.

Su, Y. and L. F. Han (1998). A new literature growth model: Variable exponential growth law of literature. *Scientometrics*, 42(2): 259-265.

Tseng, Y. H., C. J. Lin, et al. (2007). Text mining techniques for patent analysis. *Information Processing & Management*, 43(5): 1216-1247.

Vanhaverbeke, W., V. Gilsing, et al. (2012). Competence and Governance in Strategic Collaboration: The Differential Effect of Network Structure on the Creation of Core and Noncore Technology. *Journal of Product Innovation Management*, 29(5): 784-802.

Wagner, C. S. and L. Leydesdorff (2005). Network structure, self-organization, and the growth of international collaboration in science. *Research Policy*, 34(10): 1608-1618.

Waltman, L. and N. J. van Eck (2007). Some comments on the question whether co-occurrence data should be normalized. *Journal of the American Society for Information Science and Technology*, 58(11): 1701-1703.

Wang, X. W., X. Zhang, et al. (2011). Patent co-citation networks of Fortune 500 companies. *Scientometrics*, 88(3): 761-770.

Wouters, P. and L. Leydesdorff (1994). Has price dream come true - is scientometrics a hard science. *Scientometrics* ,31(2): 193-222.

WEB BASED IMPACT MEASURES FOR INSTITUTIONAL REPOSITORIES

Alastair G Smith

[1] *alastair.smith@vuw.ac.nz*
Victoria University of Wellington/ Te Whare Wānanga o te Ūpoko o te Ika a Māui, School of Information Management, Rutherford House, 23 Lambton Quay, Wellington 6140, (New Zealand/Aotearoa)

Abstract
This study investigated webometric measures that could be used to evaluate the impact of institutional repositories, using Australasian university repositories as a case study. URL citation inlinks (occurrences of the repositories' URL in the text of web pages), as found through Google searches, were counted. As well as links from the general web, links made from other Australasian academic institutions and from Wikipedia were counted. For repositories with significant deposit ratios, there appeared to be a small correlation between the URL citation inlinks from other Australasian academic institutions, and some conventional measures of research impact: the ISI citations/paper, the QS ranking score, and the ERA quality score. Repositories with higher deposit ratios appeared to achieve more inlinks from other Australasian academic institutions, indicating the value of repositories encouraging high deposit rates of their institutions research output. Institutions with repositories that had a high Wikipedia Web Impact Factor were not necessarily highly ranked in terms of inlinks from other tertiary institutions or ISI citations per paper. This indicates that repositories impact on the general web is different from their impact on the research community.

Conference Topic
Webometrics (Topic 7); Scientometric Indicators (Topic 1)

Introduction
Institutional Repositories are a common way for institutions to make their research outputs available. This study investigated opportunities for webometric evaluation of the research done by institutions through their repositories.

Most commonly webometric studies use inlink counts: links made from other websites to the site being studied. These are viewed as being analogous to citations in conventional publishing (Gairin, 1997). Either total inlink counts are used, or a Web impact factor (Almind & Ingwersen, 1997), analogous to the Journal Impact Factor for conventional publications. The Web Impact Factor is defined as the ratio of the inlink counts to a measure of the size of website, for example the number of pages.

There are a number of ways reported in the literature for counting the inlinks made to a website such as an institutional repository. Unfortunately several tools used in the past are no longer available. An early study of Web Impact Factors for Australasian universities (Smith & Thelwall, 2002) used Alta Vista to identify pages linking to the universities, but this tool is no longer available. Yahoo Site Explorer was used by many studies, for example to create a ranked list of world class universities (Ortega & Aguillo, 2009), but since Yahoo has been merged into Bing, the Site Explorer tool no longer provides useful link data. Thelwall and Sud (2011) reviewed alternative methods of estimating the online impact of organisations, including URL citation inlinks (discussed below) and organisational title mentions.

The current study used the technique of URL citation inlinks, proposed by Kousha and Thelwall (Kousha & Thelwall, 2007). This uses a search engine such as Google to locate in-text mentions of URLs associated with an institutional repository, which can be assumed to be links to documents at the repository.

There have been other webometric studies of institutional repositories. Placing research materials in repositories was found to increase the amount of data available for bibliometric analysis (Scholze, 2007). Zuccala *et al* (2007) studied an institutional repository by using web link analysis and server logs in order to investigate how users located and used the repository. An analysis of the web presence of Indian state universities (Shukla & Poluru, 2012) found that open access institutional repositories were helpful in increasing the visibility of institutions on the Web. A range of webometric measures have been used to create the Ranking Web of World Repositories (http://repositories.webometrics.info) (Aguilllo *et al* 2010) in order to support the use of repositories for scientific evaluation purposes.

A previous paper (Smith, 2012) found little correlation between impact measures of institutional repositories calculated from a new search engine Blekko (http://blekko.com/), and conventional research impact measures. The paper suggested that since links to institutional repositories are different in nature from conventional measures of research impact, measures for institutional repositories should be considered to be complementary to conventional measures, rather than directly comparable.

In the current study, impact measures were calculated based on:

- Links from the general Web
- Links from academic domains, which might be considered to be more equivalent to conventional measures of research impact
- Links from Wikipedia, which might be considered to be more indicative of the impact that the repository has in making research available to the lay community. A previous paper (Smith, 2011) identified a significant

number of links from Wikipedia to institutional repositories, and proposed that the value of institutional repositories may lie in making research available to the general Web community, rather than to the research community.

Research Questions

This study addressed the following questions:

1. What impact factor measures are appropriate for evaluating the impact of institutional repositories on the Web?
2. Do web based impact factors for institutional repositories correlate with conventional impact measures of the research output of institutions?
3. Do institutional repositories have a greater impact if a higher proportion of their research output is in the institutional repository?
4. Are there specific impact factors that reflect the different impact that institutional repositories have?

Methodology

This study investigated institutional repositories at tertiary institutions in Australasia (Australia and New Zealand). These were identified from ROAR (http://roar.eprints.org/). Repositories with less than 1000 items reported in ROAR were excluded, resulting in 39 institutional repositories being included in the study.

Google was searched with a formulation that identified pages that contained URL citation inlinks. The search excluded links from the institution itself (on the argument that these would be likely to be navigational links or self citations). Google was set not to record search history or use previous searches in interpreting the search formulation. This is important, since Google by default attempts to optimise the search on the basis of a users previous searching, which of course is counterproductive for webometric work. The data was collected in January 2013.

A typical formulation, for example for University of Auckland, was:

```
"researchspace.auckland.ac.nz" -site:auckland.ac.nz
```

This searched for web pages which included, in the text of the page, the basic URL of the archive, and excluded pages on the University of Auckland site.

Initially, only links from the institutional repository itself were excluded, for example:

```
"researchspace.auckland.ac.nz" -site:researchspace.auckland.ac.nz
```

However scanning the results indicated that this still included many pages at the institution which were either navigational in nature, or self citations (a staff member linking to their publications from their home page, for example), and it

was decided that the formulation excluding all links from the institution was more appropriate.

In some cases an institutional repository had more than one URL, for example Australian National University required a formulation:

```
"digitalcollections.anu.edu.au" OR
"dspace.anu.edu.au" -site:anu.edu.au
```

This searched for web pages that included, in the text of the page, either of the basic repository URLs, but excluded pages at the ANU site.

The Web Impact Factor for each repository was calculated by dividing the URL citation inlinks count by the number of documents in the institutional repository. The number of documents in the institutional repository was taken from ROAR.

In addition to the basic URL citation inlinks count, some counts were done of inlinks from specific types of domains.

An *Academic Institution Inlinks Count*, which might be more comparable to the citations made between research publications, was found by adding to the basic formulation a requirement that linking pages were in the Australasian academic domains, edu.au or ac.nz. So for example the formulation for University of Auckland became:

```
"researchspace.auckland.ac.nz" -site:auckland.ac.nz site:edu.au OR
site:ac.nz
```

This of course is only identifying links from Australasian academic institutions, and a more global formulation would include all academic domains (.edu, .ac.uk, individual domains of European academic institutions which generally have second level domain, for example uni-muenchen.de, etc). However this would lead to an excessively complex search formulation and it was decided that links within Australasia would give sufficient indication of the viability of the concept of an educational inlinks count.

An Academic Inlinks Web Impact Factor was calculated from the academic inlinks count, by dividing the academic institution inlinks count by the number of documents at the repository.

To measure the impact of the institutional repositories on the general lay community, a *Wikipedia inlink count* was calculated by adding to the search formulation a requirement that links were made from Wikipedia. So for example the formulation for University of Auckland became:

```
"researchspace.auckland.ac.nz"
-site:auckland.ac.nz site:wikipedia.org
```

A Wikipedia Web Impact Factor was calculated from this by dividing the Wikipedia inlinks count by the number of documents at the repository. Due to the relatively small number of Wikipedia inlinks in relation to the number of

documents in the repository, the Wikipedia Web Impact Factor was multiplied by 1000 to give a whole number.

Several conventional measures of research impact were identified and used as comparisons with the impact measures calculated in the study. These were:

- The number of citations/paper for each institution, taken from Thomson/ISI's *Essential Science Indicators*, part of the Web of Knowledge. The version used covered documents indexed by ISI in 2002-2012.
- The overall score from the QS world rankings of universities (http://www.topuniversities.com/university-rankings/world-university-rankings/2011). The QS rankings are a widely accepted measure of the quality of academic institutions worldwide. Only 26 institutions in the current study had QS ranking scores, since only the top 400 institutions worldwide were published.
- An average research excellence score derived (Hare & Trounson, 2012) from the 2010 ERA (Excellence in Research for Australia) research assessment carried out by the Australian Research Council. This of course was only available for the 29 Australian institutions in the study.

The study also looked at the extent to which an institutional repository contained a significant proportion of the research output of the institution. Mustatea (2008) found that many institutional repositories only contain a small proportion of the institution's publications when compared with the publications indexed in the ISI databases. In the current study, a ratio, the *Deposit Ratio*, was calculated. This was the ratio of documents deposited in the institutional repository, compared with the number of papers indexed by ISI in Essential Science Indicators. This is of course a crude measure, since there will be documents in the repository that would be not appropriate for indexing by ISI, and outputs indexed by ISI that may not be deposited in the repository for copyright or other reasons.

Results

The study addressed the first research question, "What impact factor measures are appropriate for evaluating the impact of institutional repositories on the Web?" by investigating a range of measures based on URL citation inlink counts to the repositories. The usefulness of these measures is addressed in the answers to the following research questions.

The second research question asked "Do web based impact factors for institutional repositories correlate with conventional impact measures of the research output of institutions?" No correlation was found between the different Web Impact Factors and the conventional measures of research impact. While it is disappointing that the Web Impact Factor of institutional repositories appears not to be a useful substitute for conventional measures of research impact, it is not

surprising. As mentioned earlier, links to institutional repositories come from different sources, and are made for different reasons, than the academic citations on which conventional measures are partly based. Also, the documents in a repository may include materials not representative of the institution's research output, including for example student work and digitised material such as historical photographs. So a Web Impact Factor calculated on the basis of the raw number of documents in the repository may not be a good measure of the inlinks per research output.

When total inlink count is considered, the picture changes a little. For the group of repositories as a whole there is no appreciable correlation between the total inlink count and the conventional measures. However if only repositories that had an Deposit Ratio of more than 1 (the ratio of the number of documents in the repository to the number of papers indexed by ISI was greater than 1) were considered, there appeared to be small but positive correlations between the total link count from educational institutions and the conventional measures: the ISI citations/paper, the QS score, and the ERA score.

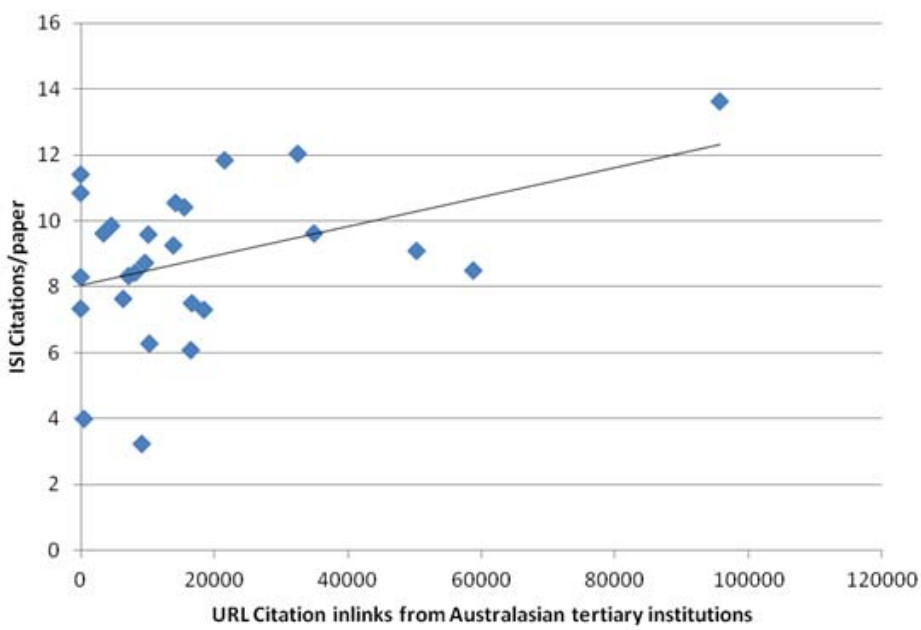


Figure 1. URL citation inlinks from Australasian tertiary institutions (Deposit Ratio>1) compared with ISI citations per paper (Pearson correlation coefficient 0.15).

Given the small numbers, the best indication of the relatively weak correlations are the scattergraph representations in Figures 1-3. For reference, the Pearson

correlation coefficients are also included. Note that QS and ERA scores were not available for all institutions. Only repositories with a Deposit Ratio greater than 1 are included.

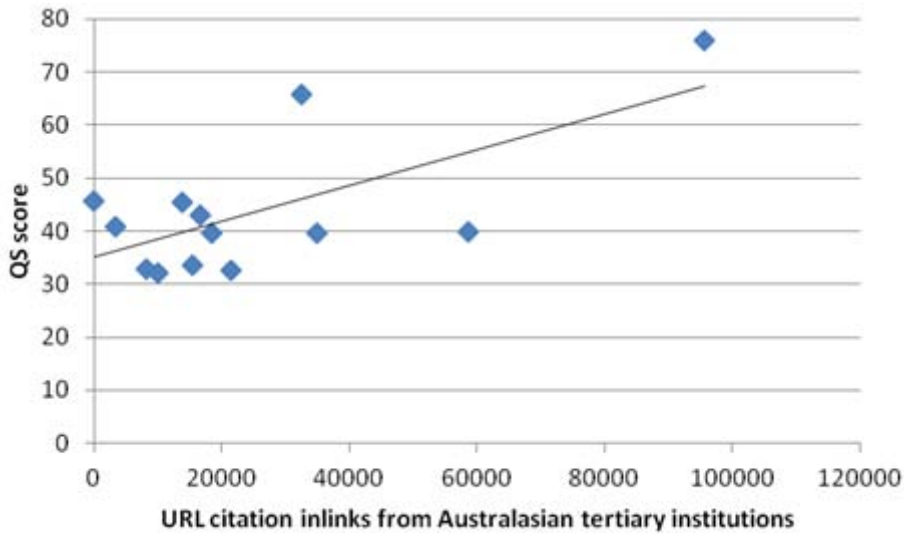


Figure 2. URL citation inlinks from Australasian tertiary institutions (Deposit Ratio>1) compared with QS score (Pearson correlation coefficient 0.14).

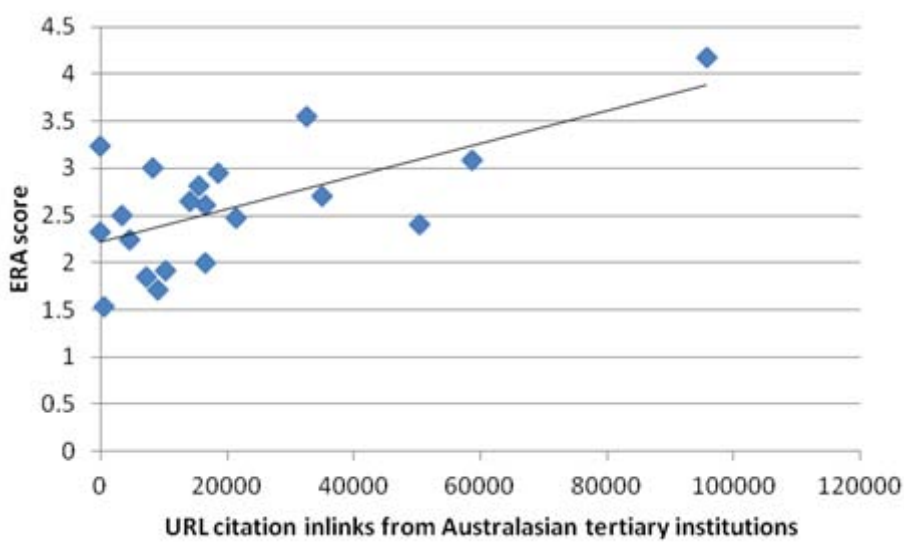


Figure 3. URL citation inlinks from Australasian tertiary institutions (Deposit Ratio>1) compared with ERA score (Pearson correlation coefficient 0.27).

This indicates that there may be value in calculating inlink counts that come from other research institutions, since these are likely to reflect the research value of the material in the repository. However the weak correlation means that further research using a larger set of repositories is needed, and that inlink counts for the institutional repository are unlikely to be a substitute for conventional measures of the research impact of institution as a whole.

Addressing the third research question, "Do institutional repositories have a greater impact if a higher proportion of their research output is in the institutional repository?" the study compared the Deposit Ratio of the repositories with URL citation inlink counts. This appeared to show a positive correlation. An indication of the relationship is shown graphically in Figure 4.

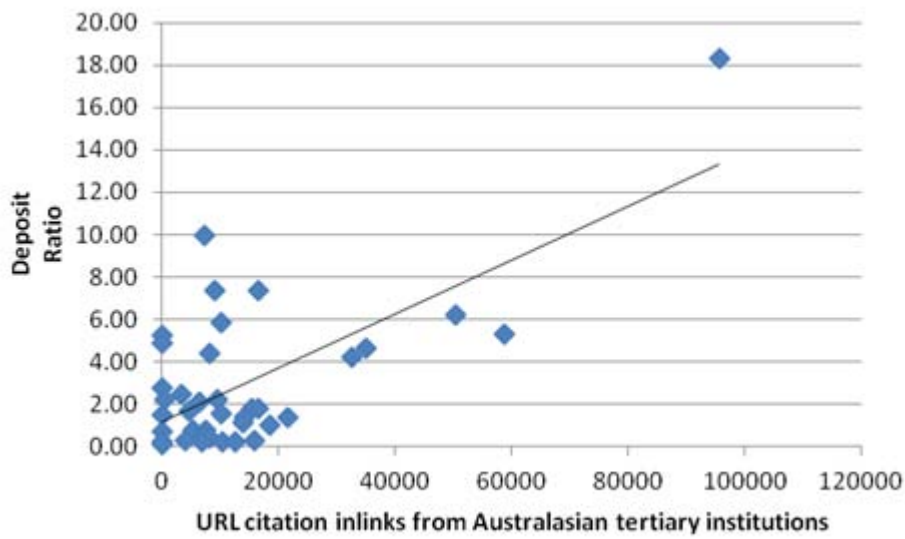


Figure 4. URL citation inlinks from Australasian tertiary institutions compared with Deposit Ratio of repository (Pearson correlation 0.68).

If this correlation is real, it appears to indicate that there is value in an institution maximising the number of research outputs in its repository, for example through mandatory deposit of publications. Higher deposit rates increase the visibility of the repository, and the chances of links being made from other research institutions.

In addressing the fourth research question, "Are there specific impact factors that reflect the different impact that institutional repositories have?" the study looked at the URL citation inlink count for links coming from Wikipedia, as a way of gauging the impact of the repositories on the general lay Web community. Neither

the Wikipedia inlink count, nor the corresponding Wikipedia Web Impact Factor, correlated with the conventional measures of research impact. This is to be expected, since Wikipedia has a different purpose than research publishing. However the Wikipedia inlink count and the Wikipedia Web impact factor provide measures of the extent to which the repository is having an impact on the general lay Web community. Table 1 shows the top 10 institutional repositories by Wikipedia Web Impact Factor (multiplied by 1000 to bring the figure to an integral number). The Web Impact Factor has been chosen in this case because Wikipedia is likely to reference many of kinds of documents that are held in institutional repositories, for example photographs. For comparison, the table also shows these institutions' ranks by the inlink count from Australasian academic institutions and by their ISI Citations per paper.

Table 1. Top 10 institutional repositories by Wikipedia Web Impact Factor.

Institution	*Wikipedia WIF (x1000)*	*Academic Citation Inlink Rank*	*ISI Citation/paper Rank*
1. University of Sydney	60	14	5
2. University of Waikato	19	17	21
3. Victoria University of Wellington	15	24	27
4. Bond University	13	16	36
5. Flinders University	12	25	16
6. University of Technology Sydney	10	7	35
7. Massey University	9	15	23
8. University of Otago	5	30	4
9. University of Tasmania	5	11	15
10. University of Canterbury	5	13	22

This indicates that institutions that have repositories with a significant impact on the general Web may not be those that have high impact in the research community.

In this study, links from Wikipedia were investigated, but of course links from other kinds of Web sites could be measures of the impact of a repository on the general Web. For example, links from blogs, Twitter feeds, Facebook, etc could be investigated.

Conclusions

This exploratory study has investigated a range of webometric measures to evaluate the impact of institutional repositories. This is important given the resources that many research institutions are investing in their repositories.

It appears that the conventional web impact factor of institutional repositories does not correlate with conventional measures of research impact. This may be

due to the number of documents in a repository not corresponding well with the conventional research output of an institution, as well as links being made to repositories for different reasons than citations are made to conventional publications. However there appears to be a small correlation between the number of links made to repositories from other academic institutions, and some conventional measures of research impact, for repositories with a high deposit ratio. This indicates that webometric measures based on links from research institutions to institutional repositories could be useful evaluation tools; particularly if the repositories achieve high deposit rates of the institutions research output. This also indicates that there may be scope for specialist web crawlers, such as the University of Wolverhampton's SocSciBot (http://socscibot.wlv.ac.uk/), to evaluate repositories for their research impact, since it appears that measures based on links from other research institutions, rather than the general web, are most valuable in terms of evaluating the research impact of the repository. There may also be value in structuring repositories in such a way that research material is differentiated from other material such as student work and digitised images.

The study also indicates that the research impact of a repository, as measured by the inlink counts from other tertiary institutions, may be enhanced by high deposit rates. While this is not surprising, it is a useful indicator to institutions that it is beneficial to encourage researchers to deposit their work in the repository.

The study also investigated measures of the repositories' impact on the general web. The specific example of links from Wikipedia was explored, showing that institutions whose repositories had a high impact in terms of their Wikipedia Web Impact Factor were not necessarily those that had a high conventional research impact. This reflects institutional repositories' value as a way of making research available to the general Web community, as well as to the research community.

This exploratory study is limited by being carried out on a limited number of institutions in a specific geographic area. Future research should see if the findings can be replicated over a broader sample of repositories. There is also scope for studies of the repositories' impact on the general Web, looking at websites such as blogs, Twitter and Facebook.

References

Aguillo, I., Ortega, J., Fernández, M., & Utrilla, A. (2010). Indicators for a webometric ranking of open access repositories. *Scientometrics*, 82(3), 477-486.

Almind, T. C., & Ingwersen, P. (1997). Informetric analyses on the world wide web: Methodological approaches to "Webometrics". *Journal of Documentation*, 53(4), 404–426.

Gairin, J. R. (1997). Valoracion del impacto de la informacion en internet: Altavista, el "citation index" de la red. impact assessment of information in the internet: Altavista, the citation index of the web. *Revista Espanola De Documentacion Cientifica*, 20(2), 175-81.

Hare, J., & Trounson, A. (2012). Excellence in research for Australia lays bare research myths. *The Australian*, (4/12/2012) Retrieved January 18 2013 from: http://www.theaustralian.com.au/higher-education/excellence-in-research-for-australia-lays-bare-research-myths/story-e6frgcjx-1225998312883

Kousha, K., & Thelwall, M. (2007). Google scholar citations and Google Web/URL citations: A multi-discipline exploratory analysis. *Journal of the American Society for Information Science & Technology*, 58(7), 1055-1065.

Mustatea, N. (2008). *To what extent is material in institutional repository representative of an institution's research output?* Report submitted to the School Of Information Management, Victoria University of Wellington in partial fulfilment of the requirements for the degree of Master of Library And Information Studies.

Ortega, J. L., & Aguillo, I. F. (2009). Mapping world-class universities on the web. *Information Processing & Management*, 45(2), 272-279.

Scholze, F. (2007). Measuring research impact in an open access environment. *Liber Quarterly: The Journal of European Research Libraries*, 17(1-4), 220-232.

Shukla, S. H., & Poluru, L. (2012). Webometric analysis and indicators of selected Indian state universities. *Information Studies*, 18(2), 79-104.

Smith, A. G. (2011). Wikipedia and institutional repositories: An academic symbiosis? *Proceedings of the ISSI 2011 Conference*, Durban, South Africa. (pp. 794-800)

Smith, A. G. (2012). Webometric evaluation of institutional repositories. *Proceedings of the 8th International Conference on Webometrics Informetrics and Scientometrics (WIS) & 13th COLLNET 2012 Meeting, Seoul, Korea*. (pp. 722-729).

Smith, A. G., & Thelwall, M. (2002). Web impact factors for Australasian universities. *Scientometrics*, 54(3), 363–380.

Thelwall, M., & Sud, P. (2011). A comparison of methods for collecting web citation data for academic organizations. *Journal of the American Society for Information Science and Technology, 62*(8), 1488-1497.

Zuccala, A., Thelwall, M., Oppenheim, C., & Dhiensa, R. (2007). Web intelligence analyses of digital libraries: A case study of the National Electronic Library For Health (NeLH). *Journal of Documentation*, 63(4), 558-89.

WHAT IS THE IMPACT OF SCALE AND SPECIALIZATION ON THE RESEARCH EFFICIENCY OF EUROPEAN UNIVERSITIES?[177]

Andrea Bonaccorsi[1], Cinzia Daraio[2*] and Léopold Simar[3]

[1]*a.bonaccorsi@gmail.com*
Dipartimento di Ingegneria dell'Energia dei Sistemi del Territorio e delle Costruzioni, University of Pisa (Italy)

[2]*daraio@dis.uniroma1.it*
* Corresponding author.
Department of Computer, Control and Management Engineering Antonio Ruberti, University of Rome "La Sapienza", via Ariosto, 25 I-00185 Roma (Italy)

[3]*leopold.simar@uclouvain.be*
ISBA, Université Catholique de Louvain, Louvain-la-Neuve (Belgium)

Abstract
The main objective of this paper is to analyse the impact of size and specialization on the research efficiency of European universities. The proposed approach builds on the notion that university production is a multi-input multi-output process different than standard production activity (Bonaccorsi and Daraio, 2004). We apply a conditional efficiency analysis approach (Badin, Daraio and Simar, 2012a,b; Daraio and Simar, 2007) in a directional distance framework to evaluate the impact of size and specialization on the research efficiency of 401 European universities, from 19 European countries. Data refer mainly to the year 2008 and include universities that in 2005-2009 have published at least 100 publications in Scopus database.
In particular we assess the impact of scale and specialization distinguishing their role on the efficient frontier and on the distribution of inefficiencies. Size seems to have a negative impact on most efficient units and on units that are lagging behind. On the contrary, specialization seems to have a slightly positive impact on the best performers and on catching-up universities.

Keywords
research evaluation, quantitative approach, scientometric indicators, European universities.

Conference Topics
Science Policy and Research Evaluation: Quantitative and Qualitative Approaches (Topic 3) and Scientometrics Indicators (Topic 1).

[177] L. Simar acknowledges research support by IAP Research Network P7/06 of the Belgian State (Belgian Science Policy).

Introduction and policy relevance

Economies of scale and scope in academic activities are the object of a new interest.

The interest in the issue of economies of scale has a number of motivations. From a policy point of view, it is important to determine whether larger units are more efficient in order to allocate public funding departing from a pure proportional formula. In some cases this has led to an explicit policy of consolidation of universities. The most known case was the Australian government decision in the late 1980s to define a minimum threshold of student size (in the range 5000-8000 students) in order to force universities to merge. More recently the Swedish government has promoted a policy of consolidation. In almost all cases the underlying assumption has been that small universities are inefficient.

From the administrator point of view the issue of economies of scale is also relevant, however. Small universities may be interested in understanding whether their size is financially sustainable in the long run, and those universities that aim to grow may want to know at which size the increases in efficiency are exhausted.

It is not surprising that a large literature has been developed on the issue of economies of scale. Brinkman and Leslie (1986) review the first 60 years of empirical studies, most of which from United States. After almost 20 years, Cohn and Cooper (2004) have offered a comprehensive survey of findings from the cost function perspective, while Johnes (2004) has reviewed the efficiency literature.

In turn, economies of scope are also discussed at length. One of the interests is in identifying and measuring the complementarity between teaching and research, which is at the core of the Humboldtian model of university (Schimank and Winnes, 2000).

Economies of scope emerge from the joint use of common inputs (Baumol, Panzar and Willing, 1982). In the case of universities, the common input is the human capital of professors and researchers. It is important to ask whether the cost of producing separately teaching and research would be lower than producing them together, keeping quality constant. The overwhelming evidence is that it is more efficient to organize teaching and research in the same organization, asking academic staff to allocate their time budget accordingly (Johnes, 2004). For example, Dundar and Lewis (1995) found that joint production is more efficient up to 300% of mean output in a sample of US universities, due to joint utilization of faculty, administrators, support staff, equipment and services. Longlong et al. (2009) argued that a reform of the Chinese system that forced researchers in the traditional Academy system to teach would generate a large increase in efficiency due to pervasive economies of scope at all levels of output.

Bonaccorsi, Daraio and Simar (2013), illustrate in details how to measure the impact of scale and scope in a directional distance framework, extending the approach of Simar and Vahnems (2012). A directional distance framework is more flexible with respect to the traditional radial approach, because it allows to choose the direction along which to assess the distance from the efficient frontier

and allows to include, in an easy way, non-discretionary inputs/outputs in the analysis.

The main objective of this paper is to analyse the impact of size and specialization on the research efficiency (given their level of teaching) of European universities that in 2009 have published at least 100 publications (including any type of documents, that is, articles, reviews, short reviews, letters, conference papers, etc) in Scopus database.
In this paper we apply the approach of Bonaccorsi, Daraio and Simar (2013) that builds on the notion that university production is a multi-input multi-output process in which, differently from standard production activity, the relationship between inputs and outputs is not deterministic (Bonaccorsi and Daraio, 2004).
In particular we assess the impact of scale and specialization separately and then jointly, distinguishing their role on the efficient frontier and on the distribution of inefficiencies.

Methodology

From a methodological point of view, there have been two major research directions.
The first has worked directly with cost functions as the dual of production functions. Here the main difficulty has been the modelling of a production function which is, by definition, not only multi-input (as any production function), but also multi-output. The traditional econometric techniques used to estimate economies of scale in a monoproduct setting were clearly inadequate.
The turning point has been the development of a full scale theory of multiproduct firm by Baumol, Panzar and Willig (1982), who introduced the distinction between ray economies of scale (long run average costs decrease with the increase in the volume of production of all outputs, keeping the proportion between various outputs constant) and product-specific economies of scale (average costs for each product decrease with the increase in the volume of that specific output). Another development was the rigorous definition of economies of scope. Based on this theory, several solutions to the problem of econometric specification were suggested, with the flexible fixed quadratic cost function (FFQC), the translog and the constant elasticity of scale (CES) as the most adopted solutions. In this line of research the existence and magnitude of economies of scale and scope is derived from the sign and size of the coefficients directly estimated on cost data.
The second has adopted the approach introduced by Farrell (1957), based on frontier analysis techniques. In this line of research most studies applying a nonparametric approach (see e.g. Grosskopf and Yaisawarng, 1990) have followed the approach developed by Fare (1986) by which, the existence and magnitude of economies of scale and scope is derived from the difference between the efficiency scores of observed DMUs and the scores that would be obtained if the specialized DMUs were aggregated. However, also this approach suffers from some weaknesses. Firstly, it introduces a sample size bias because

the number of DMUs is artificially increased, secondly it is sensitive to extremes or outliers in the data; and thirdly it always relies on the convexity assumption because data envelopment analysis is applied.

In this paper we apply a more general approach (Bonaccorsi, Daraio and Simar, 2013) to investigate on the existence of economies of scale and scope. It is based on a directional distance approach, its probabilistic characterization and uses nonparametric, nonconvex and robust to outliers efficiency estimators for the investigation of the impact of scale and scope as external environmental conditions.

Efficiency analysis techniques rely on the basic and intuitive idea of efficiency as the best use of resources (i.e. use of the lowest levels of inputs, x) to produce the maximum feasible amount of outputs (y). Related to efficiency is the concept of dominance that consists in using no more inputs to produce at least the same level of outputs and in doing better in at least one dimension.
In particular, technical efficiency can be operationalized in terms of input or output distance functions or can be measured with respect to a specific direction. The distance of each unit is measured with respect to the frontier of the production possibility set, Ψ, defined as:

$$\Psi = \{(x, y) \in R^{p+q} | x \text{ can produce } y\}. \quad (1)$$

The popular Farrell (1957) output distance of the unit (x,y) from the frontier of ψ is given by:

$$\lambda(x, y) = \sup\{\lambda > 0 | (x, \lambda y) \in \Psi\}, \quad (2)$$

and it measures the maximum feasible proportionate expansion of all outputs (y) attainable given the inputs level used (x).
Directional distances have been introduced by Chambers, Chung and Fare (1998) and are discussed at length in e.g. Fare and Grosskopf (2000). They are a generalization of the Farrell's approach. The objective of directional distances is to look for improvements in approaching the frontier in a given direction $d = (d_x, d_y)$.
A directional function, that we name also gap function, g, can be defined as:

$$g(x, y; \Psi, d_x, d_y) = \sup\{g>0 | (x\text{-}gd_x, y+gd_y) \in \Psi\}. \quad (3)$$

As it can be seen by its definition, the directional or gap function g is "additive" because it gives the amount or "gap" that has to be subtracted from the input x and at the same time has to be added to the output y in the units of the direction d to reach the frontier. On the contrary, the traditional Farrell output oriented distance function is multiplicative and can be obtained as a special case from the

directional distance, by choosing as direction d=(0,y), that is to select units own outputs as the direction vector.
It is immediate to note that the directional efficiency g, in this last specific case corresponds to the Farrell output efficiency score as follows:

$$g(x, y; \Psi, d_x = 0, d_y = y) = \lambda(x,y)\text{-}1. \quad (4)$$

If a unit has a Farrell efficiency score $\lambda(x,y)=1.2$, this means that its gap or directional efficiency score will be g=0.2 and this means that the unit has a gap of 20% in its output production: it can increase the production of its outputs by 20%. Note that the gap fuction expresses the possible improvements in units given by d_y that in our case here corresponds to the own outputs of the analysed unit.
Although in the traditional output oriented approach $\lambda \geq 1$ and $\lambda = 1$ corresponds to points that are on the efficient frontier; in the directional distance framework, $g \geq 0$ and a unit that is on the efficient frontier has a $g = 0$.
Having introduced the framework, we can now reformulate the setting in a probabilistic way. Following Daraio and Simar (2005), the joint probability measure of (X,Y) and the associated probability of being dominated, $H_{XY}(.)$ can be defined as:

$$H_{XY}(x, y) = \text{Prob}(X \leq x, Y \geq y), \quad (5)$$

and Ψ is the support of (X,Y), i.e.:

$$\Psi = \{(x, y) \in R^{p+q} | H_{XY}(x, y) > 0\}. \quad (6)$$

In this framework, Simar and Vanhems (2012) define a probabilistic version of a directional distance as follows:

$$g(x, y; \Psi, d_x, d_y) = \sup\{g>0 | (x\text{-}gd_x, y+gd_y) \in \Psi\} = \quad (7)$$
$$= \sup\{g>0 | H_{XY}(x\text{-}gd_x, y+gd_y) > 0\}.$$

A consistent nonparametric estimator of $g(x,y;\Psi,d_x,d_y)$ can be obtained by plugging a consistent nonparametric estimator of $H_{XY}(.)$ in equation (7). For further details, see Bonaccorsi, Daraio and Simar (2013).

In the lines of Daraio and Simar (2005), Simar and Vanhems (2012) introduce conditional directional distances as follows. Be $Z \in R^r$ an external or environmental variables set that might influence the production process without being inputs or outputs under the control of the unit. The conditional directional distance efficiency score g(.|z) measures the gap efficiency score given or conditioned by the external or environmental factors Z, and can be defined as follows:

$$g(x, y; \Psi, d_x, d_y | Z = z) = \sup\{g>0 | H_{XY|Z}(x\text{-}gd_x, y+gd_y | Z=z)>0\}. \quad (8)$$

A consistent nonparametric estimator of g(.|z) can be obtained by plagging a consistent nonparametric estimator of $H_{XY|Z}(.)$ in equation (8). For further details, see Bonaccorsi, Daraio and Simar (2013).

Accordingly, also robust versions of these conditional directional distances can be defined being less influenced by extremes or outliers, namely directional distance of order-m or order-α. Following Daraio and Simar (2005, 2007) the comparison of conditional efficiency scores, i.e. efficiency scores computed taking into account the external factors Z, with unconditional efficiency scores (the efficiency scores computed without taking into account the Z factors) is important to shed lights on the influence of external or environmental variables on the performance of the analysed units. In particular, in the lines of Daraio and Simar (2005; 2007), the investigation of the ratios between conditional and unconditional directional efficiency scores is relevant to assess the impact of Z on the production process of the analysed units. We define δ(Z) as the following ratio:

$$\delta(Z) = g(x, y; \Psi, d_x, d_y | Z = z) / g(x, y; \Psi, d_x, d_y) \quad (9)$$

and in the following we indicate its robust version as $\delta_\alpha(Z)$[178].

In this framework, an increasing trend (increasing regression line) of the δs with Z indicates a positive impact of the external factor (Z), whilst a decreasing trend (decreasing regression line) of δs with Z points to a negative impact of Z on the production process. A straight nonparametric regression line indicates no effect of Z on the production process. This is because when Z is favourable to the production process we expect that the conditional directional distances (defined in equation 8) will be much smaller compared to the unconditional ones for small values of Z. Therefore, the ratios (defined in equation 9) will increase with Z, on average. On the contrary, when Z is detrimental to the production process, the values of the conditional directional distances will be much smaller compared to the unconditional ones for larger values of Z. For this reason, the nonparametric regression line of over Z will be decreasing.

We adopt a directional distance framework in which we assess the distance from the frontier taking a specific direction. The direction is the factor of research (FRES) given the teaching activity carried out (TODEG5) and given the level of inputs used. In this framework, we assess the impact of size and specialization on the efficient frontier and on the distribution of inefficiency. After an analysis of their impact in isolation, we investigate their joint effect.

[178] It has to be noted that when $g(x, y; \Psi, d_x, d_y) = 0$, then, by construction $g(x, y; \Psi, d_x, d_y | Z = z) = 0$, because $0 \leq g(x, y; \Psi, d_x, d_y | Z = z) \leq g(x, y; \Psi, d_x, d_y)$, and so $\delta(Z) = 1$.

Data

We exploit a large database, recently constructed by the EUMIDA Consortium under a European Commission tender, supported by DG EAC, DG RTD, and Eurostat. This database is based on official statistics produced by National Statistical Authorities in all 27 EU countries (with the exception of France and Denmark) plus Norway and Switzerland. The EUMIDA project, relying on the results of the Aquameth project (Bonaccorsi and Daraio, 2007; Daraio et al. 2011) included two data collections: Data Collection 1 (DC 1) included all higher education institutions that are active in graduate and postgraduate education (i.e. universities), but also in vocational training. Data refer to 2008, or to 2009 in some cases.

Of these institutions, 1364 are defined research active institutions: of these only 850 are also doctorate awarding. They are the object of Data Collection 2 (DC 2), for which a larger set of variables were collected.

Table 1. Definitions of inputs, outputs and conditioning factors

Input/Output/Conditioning factor	Definition
Input	
NACSTA	Number of non academic staff
ACSTAF	Number of academic staff
PEREXP	Personnel expenditures in PPS
NOPEXP	Non-personnel expenditures in PPS
FINP	Input factor or input index including: NACSTA,ACSTAF,PEREXP,NOPEXP
Output	
TODEG5	Total Degrees ISCED5
TODEG6	Total Degrees Doctorate
INTPUB	Number of published papers in Scopus (Scimago)
FRES	Factor of research or research index, including: TODEG6, INTPUB
Conditioning factors	
TOTSTUD	Proxy of size. It is given by Total Students enrolled ISCED 5+ Total Students enrolled ISCED 6
SPEC	Specialization Index of the scientific output (Scimago)

Source: Eumida DC 2 and Scimago.

We integrate the EUMIDA data, in particular the DC 2 dataset, with the Scimago data (SIR World Report 2011, period analyzed 2005-09) that include institutions that have published at least 100 scientific documents of any type during the years 2005-2009 as collected by Scopus database. From Scimago data we used in particular the number of publications in Scopus (including any type of documents, that is, articles, reviews, short reviews, letters, conference papers, and so on, called hereafter INTPUB) and the Specialization index (SPEC) of the university

that indicates the extent of thematic concentration-dispersion of an institution's scientific output; its values range between 0 to 1, indicating generalistic vs. specialized institutions respectively. This indicator is computed according to the Gini Index and in our analysis it is used as a proxy of the specialization of the university. Table 1 defines and describes the inputs, outputs and conditioning factors that are used in the following analysis.
The monetary values are expressed in purchasing power standard (PPS). The conversion was carried out by dividing the values expressed in national currency by the respective Purchasing Power Parity (Eurostat PPP_EU27 - Purchasing power parities EU27 = 1, for the education sector), for the year 2008.
The final number of universities considered in the analysis is 401 and they come from 19 European countries. We excluded in fact, universities for which expenses, or number of academic staff or number of students, or number of publications data were not available.

The following Table 2 shows some descriptive statistics on inputs, outputs and conditioning factors used in the analysis.

Table 2. Descriptive statistics on inputs, outputs and conditioning factors- whole sample (401 obs.)

	Minimum	Maximum	Mean	Std. Deviation
NACSTA	59.00	8606.00	1496.89	1408.39
ACSTAF	65.00	6571.00	1470.21	1058.13
PEREXP	4501077.76	674760008.45	142577882.82	121662901.93
NOPEXP	5104884.60	699593733.99	87111330.32	94924980.37
TODEG5	.00	28215.00	3881.57	3146.21
TODEG6	.00	1855.00	200.72	214.42
INTPUB	300.00	33610.00	5570.78	5625.99
TOTSTUD	331.00	181693.00	20258.25	17485.77
SPEC	.40	1.00	.69	.13

Results

We run a preliminary descriptive analysis on the variables reported in Table 1 and we decided, on the base of this investigation, given the very high correlations found (higher than 0.90), to aggregate the inputs (NACSTA, ACSTAF, PEREXP, NOPEXP) in a single input index, named FINP, and two of the outputs (TODEG6, INTPUB), highly correlated, in a research index, named FRES.

The model formulated for the estimation of the technical efficiency of European universities is based on a output oriented directional distance function, in which we use the input index as an input and two outputs, namely the research index and the Total number of Degrees at ISCED5 level (TODEG5), keeping this last one as a nondiscretionary output. Summing up, we estimate the technical efficiency of European universities in their research activity, proxied by the research index,

given the level of teaching they are running, as proxied by TODEG5. We estimate several nonparametric and nonconvex efficiency scores that are summarized in Table 3, where we report some descriptive statistics on the computed efficiency scores. N_DOM is the number of points that dominate the analysed unit; on average the European universities analysed in this paper, are dominated by 7 other European universities; however, an high variation exists as N_DOM goes from a minimum value of 0 to a maximum value of 95.
FDH_DIR is the output directional distance efficiency measure computed on a Free Disposal Hull estimator (Deprins, Simar and Tulkens, 1984).
GAPS_TODEG6 measures the existing gap (compared to the most efficient units) on the variable TODEG6 existing when an FDH_DIR efficiency approach have been applied. The European universities analysed in this paper, given their inputs and the level of teaching which are carrying out, could have produced on average 79 more Doctorate Degrees (TODEG6).
GAPS_INTPUB measures instead the existing gap in the production of number of papers (INTPUB); on average the European universities in our sample could produce 2114 more papers, given their inputs and their level of teaching.
Robust_DIR is the output directional distance efficiency measure computed on a robust nonparametric estimator that does not envelop the 5% of most efficient units in the sample. The following gaps reported in Table 3. Rob_GAPS_TODEG6 and Rob_GAPS_INTPUB, are the estimated gaps, in TODEG6 and in INTPUB respectively, obtained by applying Robust_DIR directional efficiency measure, and are computed taking out from the comparison the 5% of the most efficient units. It appears that the European universities analysed could produce, on average, 22 more Doctorate Degrees and could publish 578 more papers. The high standard deviation of the robust gaps and the big range of variation of gaps (from -384 to 372 for TODEG6; from -10305 to 9970 for INTPUB) points to the existence of heterogeneous results. This heterogeneity could be due also to differences in the data available for the different countries and more investigations on the comparability and reliability of data have to be carried out.

Table 3. Descriptive statistics on the efficiency analysis results (whole sample 401 obs.)

	Minimum	Maximum	Mean	Std. Deviation
N_DOM	.00	95.00	7.05	12.58
FDH_DIR	.00	9.81	.76	1.23
GAPS_TODEG6	.00	581.31	78.83	99.30
GAPS_INTPUB	.00	15588.00	2113.66	2662.61
Robust_DIR	-.49	9.08	0.53	1.21
Rob_GAPS_TODEG6	-384.33	371.82	21.54	79.73
Rob_GAPS_INTPUB	-10305.40	9969.95	577.62	2137.88

Note: robust efficiency scores calculated with alpha=0.95.

Detailed summary statistics by country are not reported to save space.

After that, we analysed the impact of size and of scientific specialization, firstly in isolation and after that jointly. Size is proxied by the total number of enrolled students at all levels (both graduate and post graduate ones), TOTSTUD, while the scientific specialization is proxied by the variable SPEC.
Considered in isolation, we observed that size seems to have no effect on the research efficiency of the analysed European universities, given their level of teaching. In particular, Size does not have any clear effect on the most efficient universities, and seems to have almost no effect on the universities that are lagging behind. A similar effect is observed for scientific specialization. It seems that SPEC does not play any role on the efficiency if considered in isolation.
The joint impact of size and specialization is illustrated in the following figures.
Figure 1 illustrates the impact of size and specialization on the efficient frontier of analysed units. In particular, the 3-dimensional plot reported on the left illustrates the $\delta(Z)$ defined in equation (9), i.e. the ratios between conditional and unconditional efficiency scores, versus the conditioning factors, Z, that in this case are size (Z_1) and specialization (Z_2). The two panels reported on the right illustrate the marginal effect of each Z variable on the efficient frontier. The top panel on the right of Figure 1 reports the impact of size on the efficient frontier; whilst the bottom panel on the right of Figure 1 presents the impact of specialization on the efficient frontier of units. Globally, it appears that there is an U-shape effect of size (Figure 1 top panel on the right): up to around 45000 enrolled students (TOTSTUD) we observe a decreasing trend of the deltas which points to a negative impact of size on the research efficiency of universities whilst, after 45000 TOTSTUD we see a small range of increasing trend (positive impact) which should be interpreted with care, because it is determined only by a few large universities. On the contrary, SPEC seems to have an homogeneous or slightly increasing trend on research efficiency (Figure 1 bottom panel on the right).
Figure 2 shows the impact of size and specialization on the distribution of inefficiencies of the analysed units. The 3-dimensional plot reported on the left illustrates the $\delta_\alpha(Z)$ that are the robust deltas computed with respect to a median frontier (to catch the impact on the average of the distribution of the inefficiencies and not on the most efficient boundary as in the previous figure) versus the conditioning factors size (Z_1) and specialization (Z_2). The top panel on the right of Figure 2 presents the impact of size on the distribution of inefficiencies; whilst the bottom panel on the right of Figure 2 shows the impact of specialization on the units that are lagging behind. Generally, it seems that there is a negative impact of size on the distribution of inefficiencies among universities (decreasing trend, see Figure 2 top panel) whilst specialization seems to have an inverted U-shape effect with a positive impact (up to 0.8) and then a negative impact, even if in the range with SPEC higher than 0.8 there are only a few observations (see Figure 2 bottom panel).

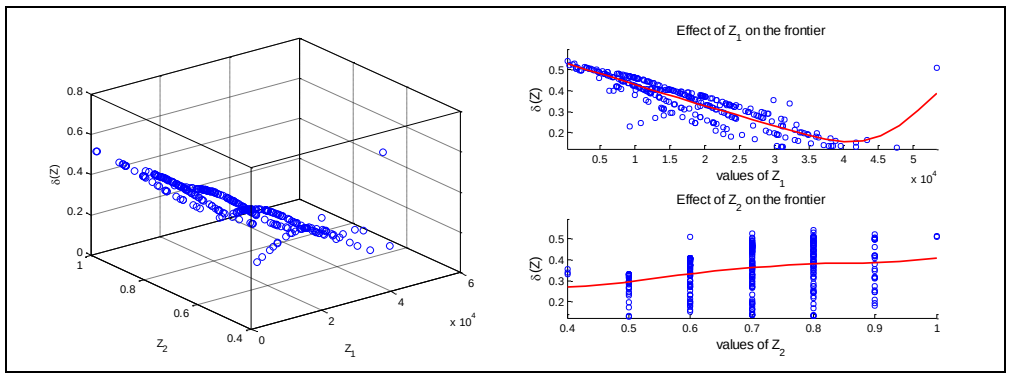


Figure 1. Joint impact of Size -TOTSTUD (Z1) and Specialization - SPEC (Z2) on the efficient frontier (of research given teaching).

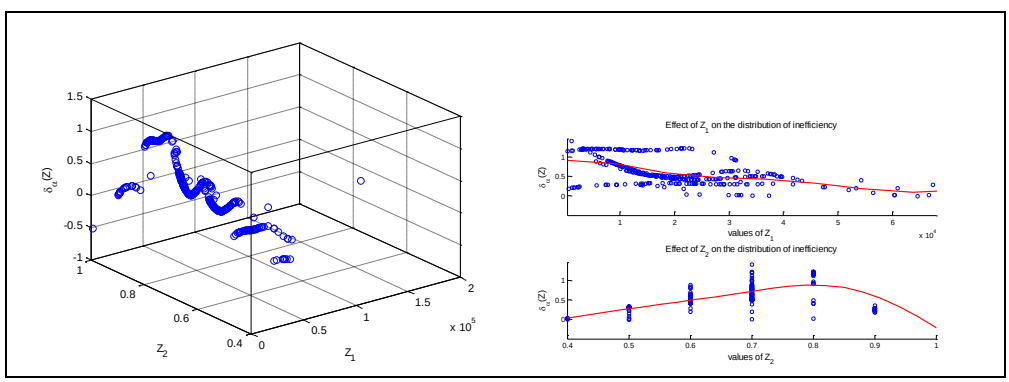


Figure 2. Joint impact of size -TOTSTUD (Z1) and Specialization - SPEC (Z2) on the distribution of inefficiency.

Conclusions

In this paper we apply a general approach (Bonaccorsi, Daraio and Simar, 2013) to investigate on the existence of economies of scale and scope in research efficiency of European universities given their level of teaching. It is based on a directional distance approach, its probabilistic characterization and uses nonparametric, nonconvex and robust to outliers efficiency estimators for the investigation of the impact of scale and scope as external environmental conditions.

We disentangled the impact of scale and specialization distinguishing their role on the efficient frontier and on the distribution of inefficiencies. We find that size seems to have a negative impact on most efficient units and on the distribution of inefficiencies, with the exception of a few large universities that remain isolated. On the contrary, specialization seems to be positive both for efficient units and for units that are lagging behind, except for a few highly specialized universities that seems to suffer from a negative impact of SPEC on the research efficiency.

The great heterogeneity in the performance found, shows that there exists large difference across European universities. This heterogeneity could also be due to differences in the data available for the different countries and more investigations on the comparability and reliability of data have to be carried out.

Selected References

Badin L., Daraio C., Simar L. (2012a), How to Measure the Impact of Environmental Factors in a Nonparametric Production Model, *European Journal of Operational Research*, 223, 818–833.

Badin L., Daraio C., Simar L. (2012b) Explaining Inefficiency in Nonparametric Production Models: the State of the Art, *Annals of Operations Research*, DOI 10.1007/s10479-012-1173-7.

Baumol, W.J., J.C. Panzar and R.D. Willig (1982), *Contestable Markets and the Theory of Industry Structure*, New York: Harcourt Brace Jovanovich.

Bonaccorsi A., Daraio C. (2004), "Econometric approaches to the analysis of productivity of R&D systems. Production functions and production frontiers", in H.F. Moed, W. Glanzel and U. Schmoch (edited by), *Handbook of Quantitative Science and Technology Research, Kluwer Academic Publishers*, 51-74.

Bonaccorsi A., Daraio C. (2005), "Exploring size and agglomeration effects on public research productivity", *Scientometrics*, Vol. 63, No. 1, 87–120.

Bonaccorsi A., Daraio C. (2007), eds, *Universities and Strategic Knowledge Creation. Specialization and Performance in Europe*, Edward Elgar Publisher, Cheltenham (UK).

Bonaccorsi, A., Daraio C., Simar, L. (2006), "Advanced indicators of productivity of universities. An application of Robust Nonparametric Methods to Italian data", *Scientometrics*, Vol. 66, No. 2, 389-410.

Bonaccorsi A., Daraio C. and Simar L. (2013), Economies of scale and scope using directional distances with Application to European universities, Technical report DIAG, Roma.

Brinkman, P.T. and L.L. Leslie (1986) "Economies of scale in higher education: Sixty years of research". *The Review of Higher Education*, 10 (1), 1-28.

Cohn E. and S.T. Cooper (2004) 'Multi-product cost functions for universities: economies of scale and scope'. In G. Johnes and J. Johnes (Eds.), *The International Handbook on the Economics of Education*. Cheltenham: Edward Elgar.

Daraio C., Simar, L. (2005), "Introducing Environmental Variables in Nonparametric Frontier Models: a Probabilistic Approach", *The Journal of Productivity Analysis*, 24 (1), pp. 93-121.

Daraio C., Simar L. (2007), *Advanced Robust and Nonparametric Methods in Efficiency Analysis. Methodology and Applications*, Springer, New York (USA).

Daraio C. et al. (2011), The European University landscape: A micro characterization based on evidence from the Aquameth project, *Research Policy*, 40, 148–164.
Dundar, H. and D.R. Lewis (1995) 'Departmental productivity in American Universities: Economies of scale and scope'. *Economics of Education Review*, 14(2), 119–144.
Fare, R, (1986) Addition and efficiency, *Quarterly Journal of Economics*, 101(4), 861-865.
Farrell, M. (1957), 'The measurement of productive efficiency', *Journal of the Royal Statistical Society*, Series A, 120, 253–81.
Grosskopf, S., S. Yaisawarng, (1990), Economies of scope in the provision of local public services, *National Tax Journal*, 43(1), 61-74.
Johnes, J. (2004). 'Efficiency measurement'. In G. Johnes, & J. Johnes (Eds.), *The International Handbook on the Economics of Education*. Cheltenham: Edward Elgar.
Longlong, H., L. Fengliang, M. Weifang (2009). 'Multi-product total cost functions for higher education: The case of Chinese research universities'. *Economics of Education Review*, 28, 505–511.
Schimank U. and M. Winnes (2000), 'Beyond Humboldt? The relationship between teaching and research in European university systems. *Science and Public Policy*, 27, 398-408.
Simar, L. and Vanhems A. (2012), Probabilistic Characterization of Directional Distances and their Robust Versions, *Journal of Econometrics*, 166, 342-354.

WHICH FACTORS HELP TO PRODUCE HIGH IMPACT RESEARCH? A COMBINED STATISTICAL MODELLING APPROACH

Fereshteh Didegah[1], Mike Thelwall[2], Paul Wilson[3]

[1]f.didegah@wlv.ac.uk, [2]m.thelwall@wlv.ac.uk, [3]pauljwilson@wlv.ac.uk
[1,2,3]Statistical Cybermetrics Research Group, School of Technology, University of Wolverhampton, Wulfruna Street, Wolverhampton WV1 1LY UK

Abstract
This study uses an appropriate statistical model to simultaneously assess five factors under the control of researchers that may help to produce highly cited research: individual collaboration; institutional collaboration; international collaboration; research funding; and abstract readability in Biology & Biochemistry, Chemistry and Social Sciences. Using a negative binomial-logit hurdle model, the results show that individual collaboration is a significant determinant of citation counts in all three areas. Institutional teamwork gives a citation advantage in Social Sciences but international teamwork shows no significant association with citation counts. Research funding very significantly associates with increased citation counts in Biology & Biochemistry and Chemistry but abstract readability is not found to be significant. In summary, individual teamwork and research funding are the keys to high impact research, at least in these three areas.

Conference Topic
Scientometrics Indicators: Introduction
- Criticism and new developments
- Relevance to Science and Technology, Social Sciences and Humanities (Topic 1).

Introduction
Conducting high impact research seems to be a common goal for researchers. For example, the link between excellent writing skills and high impact research has been extensively discussed, mainly based upon the advice of senior researchers (Zimmerman, 1989). Citation counts are widely acknowledged as the main research impact indicator and empirical studies have been carried out to seek associations between citation counts and various objective and easily measurable properties of the research. These include the impact of the publishing journal (Boyack & Klavans, 2005), research collaboration (Gazni & Didegah, 2010), the interdisciplinarity of the article references (Larivière & Gingras, 2010), the number and impact of references (Boyack & Klavans, 2005), and the size of the related field (Lovaglia, 1989).
Based on the above findings, authors seeking to maximise the impact of their research may select high impact journals to publish in and may also be particularly careful to ensure that their literature review does not miss any

relevant highly cited papers. If they wish to conduct high impact type of research then they may also seek to engage in collaborations (hence generating more co-authors). Nevertheless, unscrupulous scholars may even examine lists of significant factors associated with increased citations and try to manipulate those that they have control over. For example they may invite scholars to be co-authors when their contribution does not merit it or cite largely irrelevant high impact work.

This study examines the association between research collaboration, research funding and article abstract readability and citation counts. The goal is to assess which factors under the control of researchers are most important for the production of high impact research. Research collaboration has been frequently analysed (Sooryamoorthy, 2009) and the other two factors have also been examined (Zhao, 2010; Gazni, 2011) but they have not been examined simultaneously for multiple research fields using an appropriate statistical model. This is an important omission because inappropriate models may generate misleading conclusions and non-simultaneous tests may identify apparently important factors that are not relevant when other factors are also considered. This study fills this gap by applying a negative binomial-logit hurdle model to three different subjects.

Literature review

As introduced above, research citation impact has been shown to be related to a number of objective properties of articles. One of the most common factors positively associated with increased citations is research collaboration (Gazni & Didegah, 2010; Sooryamoorthy, 2009). Publishing in a high impact journal is also one of the foremost factors associated with higher citation counts (Boyack & Klavans, 2005). The reputation of authors is also known a determinant of citation impact (Peters & van Raan, 1994). The five factors that are examined in this study are individual collaboration, institutional collaboration, international collaboration, research funding and readability of abstracts. Among the factors previously studied, these factors seem to be most under the control of the authors. Publishing in high impact journals was not included because this is itself an indicator of successful research and is an external approval indicator, like citation counts, rather than being an aspect of the research itself. Literature on the three factors is reviewed in the next sub-sections.

Research collaboration

Multi-author research is becoming more common (Gazni, Sugimoto, & Didegah, 2012; Persson, Glänzel, & Danell, 2004) and receives more citations than solo research (Gazni & Didegah, 2010; Sooryamoorthy, 2009; Leimu & Koricheva, 2005a&b). However, a few studies have found no correlation between more authors and increased citations (Bornmann, Schier, Marx, & Daniel, 2012; Haslam et al., 2008). These studies' findings are often not generalizable, however because they are limited to a single country (Sooryamoorthy, 2009), a single

institution (Gazni & Didegah, 2010), a single field of study (Leimu & Koricheva, 2005a&b; Haslam et al., 2008) or a specific journal (Bornmann, Schier, Marx, & Daniel, 2012). Using correlation and regression tests, a correlation between citation counts and the number of authors has been found (Gazni & Didegah, 2010; Sooryamoorthy, 2009; Leimu & Koricheva, 2005a&b; Haslam et al., 2008) but the extent to which the number of authors contributes to increased or decreased citations has not been determined. The difference between the results of previous studies might have resulted from the differing samples of publications used and, in particular, there may be disciplinary differences. Whereas previous studies have conducted detailed micro-level analyses, the current study is done at a macro level and is not limited to a single country, institution, field or journal.

International collaboration can also lead to increased citation counts (Sooryamoorthy, 2009; Glänzel, 2001; Glänzel & Schubert, 2001; Katz & Hicks, 1997; Narin, Stevens, & Whitlow, 1991). Conversely, however, an investigation of Harvard University publications found no correlation between international collaboration and citation counts (Gazni & Didegah, 2010), and this may be a special case for Harvard, as a world-leading institution. Most studies are geographically or institutionally limited and hence are difficult to generalise. Two studies (Glänzel, 2001; Glänzel & Schubert, 2001) avoid this issue by taking the full Science Citation Index (SCI) during a one or two-year period. However, they do not cover social sciences fields. This research fills this gap in the literature by studying social sciences in comparison to life and physical sciences. To measure the impact of international collaboration on citation counts, the very simple method of comparing the mean citation of domestic collaboration with international collaboration is often used. This has the limitation that the difference may be spurious: caused by factors other than the ones investigated. International collaboration seems to be particularly beneficial for small institutions (Goldfinch, Dale, & DeRouen, 2003) rather than big institutions (Gazni & Didegah, 2010).

Institutional collaboration, which involves researchers from different institutions, also associates with the higher citation impact of papers (Gazni & Didegah, 2010; Sooryamoorthy, 2009; Narin & Whitlow, 1990). These studies are also geographically and institutionally limited and do not have the coverage of the current study. A simple correlation was tested to examine the association between institutional collaboration and citation counts.

Abstract readability

Readability refers to the level of difficulty of the language used to write a text. Using the Flesch difficulty score, Gazni (2011) found that papers with less readable abstracts were cited more than the papers with more readable abstracts in the five top institutions in the world. It may be that in the world's top institutions their high prestige ensures that their less readable abstracts seem more impressive, whereas unreadable abstracts may be taken as a sign of incompetence for researchers at other institutions. Alternatively, less readable abstracts may associate with higher citation areas of study, such as the more quantitative fields.

However, medical articles with structured abstracts, using different sections in a way that is known to be more readable (Hartley & Benjamin, 1998), are, on average, more cited than articles with traditional unstructured abstracts (Hartley & Sydes, 1997).
It seems that there is not a strong relationship between article readability and citation impact in the sub-fields of Social Sciences: Marketing, Psychology and Education Science (Stremersch, Verniers, & Verhoef, 2007; Hartley, Sotto, & Pennebaker, 2002; Hartley & Trueman, 1992). Finally, three decades ago, Bottle, Rennie, Russ and Sardar (1983) claimed that the readability of articles was significantly decreasing although the reasons for this were not clear and it is not known if this trend has continued.
Given that abstract readability and its association with research citation impact has been studied only to a limited degree, larger scale investigations are needed. This study partly addresses this demand.

Research funding

It is widely believed that insufficient funding can lead to shortcomings in research (Reed et al., 2007). For example, a higher citation impact is expected when funding is provided (Levitt, 2011). A number of studies have claimed an association between research impact and funding in Medical Education research (Read et al., 2007), Library and Information Science (Zhao, 2010), and Biomedical research (Lewison & Dawson, 1998) but with the caveat that it varies across subject domains in a single country (Jowkar, Didegah, & Gazni, 2011). However, a decade before Zhao (2010), Cronin and Shaw (1999) did not find an association between research grants and the citation impact of papers in Information Science. Research funding also seems not to be a significant determinant of increased citations in Psychology (Haslam et al., 2008) and so there may be disciplinary differences in the importance of funding. The researchers basically compared the average citations of the entire funded research with that average of the unfunded research in a single field whereas this study will examine and compare the citation impact of funded vs. unfunded research at a paper level.

Research questions

The factors examined here have been previously investigated: particularly research collaboration and research funding. The current study aims to fill three knowledge gaps in the literature: first, there is a lack of consensus on the influence of citation factors since different studies came to different conclusions on the effect of a specific factor; second, the literature is silent on the extent to which the factors determine the impact; and lastly, most literature on the influence of factors considered them separately and mostly within a single field. There is a particular problem with overlapping factors, such as collaboration and internationality. For example, more international papers tend to have more authors so if international research is more cited is this because it is international or

because it has more authors (and vice versa)? Therefore, this study seeks to simultaneously analyse the three factors of citation impact in three different fields of research that are representatives of three broad areas of science (Life Sciences, Physical Sciences and Social Sciences). It goes further than the simple correlation between the factors and citation impact and provides evidence of the extent to which these factors associate with increased or decreased citations. This study seeks to answer two research questions:

1. Which factors under the control of the researcher associate with increased citation impact, taking into account that some of these factors overlap? This concerns individual collaboration, institutional collaboration, international collaboration, research funding and abstract readability.
2. *To what extent* do the determinants of citation impact associate with increased citation counts?

Methods

Publications from Biology & Biochemistry, Chemistry and Social Sciences covered by Thomson Reuters' Web of Science (WoS) from 2000-2009 were extracted (16,378 articles in Chemistry, 16,058 articles in Biology & Biochemistry, and 15,932 articles in Social Science) by systematic sampling based upon the year of publication and the sub-fields. Using the list of journals provided by *ScienceWatch.com* classifying each journal into one of the 22 ESI fields, a journal-based method of searching was used to find and download the related publications. Only two types of documents, articles and conference proceedings, were included because original research is mainly published in these two types of documents (Milojević & Leydesdorff, 2012).

Although the subject classification in WoS is journal-based, it is well-established and has frequently been used by scientometricians to classify individual papers. The three fields were picked up from a list of 22 different subject fields classified by Essential Science Indicators (ESI) in WoS. Biology & Biochemistry was chosen as a representative for life sciences and Chemistry was chosen as a representative for physical sciences (See Nagaoka, Igami, Eto, & Ijichi (2011) for the categorization of subject fields), as they both are the largest fields (based on number of their publications) in their own category.

Dependent and independent variables

The number of citations to papers is the dependent variable and the independent variables are research collaboration, research funding and readability of abstract. Three different patterns of research collaboration were used: individual collaboration (number of authors in each paper), institutional collaboration (number of institutions in each paper) and international collaboration (number of countries in each paper). To measure individual collaboration, the number of authors per paper was automatically counted from the WoS authors' names field. To identify and count institutional and international collaborations, the number of distinct institutions and countries contributing to the WoS affiliation field of each

paper was automatically counted. A research paper was counted as funded if there was an entry in its WoS funding field. WoS contains funding acknowledgement data from August 2008 forward (Thomson Reuters Technical Support, 2013). Due to the data limitation, we could not include funding variable in the model for the all ten years. We ran extra models and included funding variable with the other four variables for 2009 data only. The Flesch Reading Ease Score was used to measure the readability of the abstracts. There are numerous formulae to measure the readability of a text but the Flesch score seems to be the most popular and the Microsoft Office Word 2010 API was used to automatically calculate it.

Statistical procedures

Count models provide a structural framework for analysing the count data. Given that the dependent variable of our study is count data (citation counts), these types of regression models are the most appropriate. The research data set is overdispersed (i.e. the variance of the data is greater than its mean). A Poisson regression model, the basic count model, assumes mean and variance equality (Cameron & Trivedi, 2001); therefore a Poisson model cannot adequately deal with overdispersed data and this option was rejected.

Initially, standard, zero-inflated and hurdle negative binomial models were considered. A *standard* negative binomial model is frequently used to model overdispersed data. *Hurdle models* seek first to determine the probability of an observation being positive or zero, and then the parameters of the count distribution for positive observations. *Zero-inflated* models assume two types of zeros in the data: zeros which arise from a count distribution and zeros which arise from a "perfect-zero" distribution (Hilbe, 2011). We fitted these three models on the dataset and hurdle models were found to give the best fit to the data. The hurdle model is also intuitively a good choice because it seems reasonable to assume that it is a significant hurdle for a paper to receive its first citation but after this it is more likely to be cited in the future. More citations may occur because a cited paper is listed higher in information retrieval systems (e.g., Google Scholar) or because the endorsement of a citation reported in such systems.

There are different types of hurdle model. Logit and complementary log-log (cloglog) hurdle models were fitted on the data set and found to have identical AIC values. AIC (Akaike Information Criterion) is an indicator of the statistical goodness of fit and helps to choose between two models. The logit and cloglog models are the binary models for modelling the zero counts and specify the relationship between the predictors and the dependent variable. As the results from the logit model are easier to interpret, it was decided to report the logit model (Hilbe, 2011). In the negative binomial-logit hurdle model, two parameters are predicted with the negative binomial model: The overdispersion parameter and the mean of negative binomial model. With the log model, odds ratio in the form of Log [P(citations>1)/P(citations=0)] is predicted.

Table 1. The results of hurdle model in Biology & Biochemistry (2000-2009)

Logit model	*Coef.*	*Exp(Coef.)*	*Std. Err.*	*z*	*P>z*	*[95% Conf.*	*Interval]*
No. of Authors	0.06	1.062	0.012	4.89	0.000	0.036	0.084
No. of Countries	0.122	1.129	0.06	2.04	0.042	0.005	0.239
Readability of Abs.	-0.005	0.995	0.002	-2.21	0.027	-0.01	-0.001
Constant	2.033	7.64	0.096	21.13	0.000	1.845	2.222
NB model	*Coef.*	*Exp(Coef.)*	*Std. Err.*	*z*	*P>z*	*[95% Conf.*	*Interval]*
No. of Authors	0.05	1.051	0.004	12.63	0.000	0.042	0.058
No. of Countries	0.029	1.03	0.02	1.47	0.143	-0.01	0.069
Readability of Abs.	-0.01	0.99	0.001	-10.71	0.000	-0.012	-0.008
Constant	2.558	12.91	0.034	74.78	0.000	2.491	2.625
alpha (the dispersion parameter)	0.572	1.772	0.021	26.67	0.000	0.53	0.614

Results

The results of negative binomial-logit hurdle model provide coefficients for both the negative binomial (non-zero citation counts) and the logit (proportion of uncited papers) components of the model (tables 1 to 7). For each subject category, two hurdle models are run, one for the whole ten years (2000-2009) excluding research funding variable and another only for 2009 including research funding.

Biology & Biochemistry In Biology & Biochemistry (2000-2009), the coefficients of the negative binomial model show that only the number of authors associates with increased citations. The positive significant coefficient for the number of authors indicates that a one-unit change in the number of authors increases the mean citation count by 5%: for papers that are cited at least once, each extra author attracts, on average, 5% more citations. The number of countries is not a significant determinant of citation counts ($p > 0.05$) in this field. With respect to the logit model, the number of countries is significantly associated with decreased zero citations and a one-unit change in the number of countries decreases the mean number of zero citation articles by 13% (Table 1). To handle the collinearity issue (the high correlation between the number of institutions and the two other research collaboration factors) the number of institutions was removed from the model since there is a high correlation between this variable and the number of authors and the number of countries. The effect of this variable on citation counts was separately scrutinized in more detail. Keeping the number of authors and the number of countries constant at different values, extra hurdle models were run. In the majority of cases, the coefficient of the number of citations is not significant and the results are not consistent and vary from one number of countries to another. So the overall evidence of the impact of the number of institutions in Biology & Biochemistry is unclear (Table 4), but it seems that this is not an important factor. Abstract readability is strongly associated with decreased citation counts and increased zero citations. Whilst abstract readability is highly statistically significant, the exponential coefficient of 0.99 in both the negative binomial and logit models indicates that this variable has no practical significance (Table 1).

Table 2. The results of hurdle model in Biology & Biochemistry including research funding (2009)

Logit model	*Coef.*	*Exp (Coef.)*	*Std. Err.*	*z*	*P>z*	*[95% Conf. Interval]*	
No. of Authors	0.045	1.046	0.022	2.080	0.037	0.003	0.088
No. of Countries	-0.018	0.982	0.102	-0.180	0.861	-0.218	0.182
Funding	0.658	1.931	0.121	5.450	0.000	0.421	0.895
Readability of Abs.	-0.008	0.992	0.005	-1.670	0.095	-0.017	0.001
Constant	0.217	1.242	0.199	1.090	0.276	-0.173	0.607
NB model	*Coef.*	*Exp (Coef.)*	*Std. Err.*	*z*	*P>z*	*[95% Conf. Interval]*	
No. of Authors	0.059	1.060	0.015	4.020	0.000	0.030	0.087
No. of Countries	0.070	1.072	0.079	0.880	0.378	-0.086	0.226
Funding	0.248	1.282	0.104	2.380	0.017	0.044	0.453
Readability of Abs.	-0.010	0.990	0.004	-2.720	0.007	-0.018	-0.003
Constant	0.370	1.448	0.179	2.070	0.038	0.020	0.720
alpha	0.566	1.760	0.156	3.630	0.000	0.260	0.871

According to the results of the negative binomial model in Biology & Biochemistry (2009), research funding is strongly associated with increased citations and one-unit change in the research funding increases the mean citation count by 28.2%. Other factors are behaving the same to the ten-year model. With respect to the logit model, research funding is significantly associated with decreased zero citations (Table 2).

Table 3. The results of hurdle model in Chemistry (2000-2009)

Logit model	*Coef.*	*Exp(Coef.)*	*Std. Err.*	*z*	*P>z*	*[95% Conf. Interval]*	
No. of Authors	0.076	1.079	0.012	6.24	0.000	0.052	0.1
No. of Countries	0.357	1.428	0.06	5.95	0.000	0.239	0.474
Readability of Abs.	-0.003	0.997	0.002	-2.17	0.03	-0.007	0.000
Constant	1.031	2.805	0.083	12.46	0.000	0.869	1.193
NB model	*Coef.*	*Exp(Coef.)*	*Std. Err.*	*z*	*P>z*	*[95% Conf. Interval]*	
No. of Authors	0.017	1.018	0.006	2.92	0.004	0.006	0.029
No. of Countries	0.049	1.05	0.028	1.77	0.078	-0.005	0.104
Readability of Abs.	-0.008	0.992	0.001	-8.75	0.000	-0.01	-0.006
Constant	2.165	8.718	0.044	49.62	0.000	2.08	2.251
alpha	0.869	2.384	0.03	28.71	0.000	0.81	0.928

Chemistry In Chemistry (2000-2009) with respect to the negative binomial part of the hurdle model, the number of authors is the only determinants of increased citations and a one-unit change in the number of authors increases the expected mean citation count by about 2%. With respect to the logit model, the number of authors and the number of countries are significantly associated with decreased zero citations and a one-unit change in the number of authors and the number of countries decreases the expected mean zero citation by around 8% and 43%. The number of institutions was removed from the model due to a high collinearity. However, fixing the number of authors and the number of countries at different values, we again ran extra hurdle models to more precisely study this variable but no clear evidence was obtained (Table 7). Although abstract readability is

statistically significantly associated with decreased citation counts and increased zero citations, its association is of no practical significance (Table 3).
With respect to the negative binomial model in Chemistry (2009), research funding is a significant determinant of increased citations and a one-unit change in this variable increased the mean citation counts by 42.3%. According to the logit part of the hurdle model, research funding is associated with decreased zero citations and a one-unit change in this variable decreases zero citations by 8.2%. The number of authors is not a significant determinant of increased citations in the one-year model but it is significantly associated with decreased zero citations (Table 4).

Table 4. The results of hurdle model including research funding in Chemistry (2009)

Logit model	*Coef.*	*Exp(Coef.)*	*Std. Err.*	*z*	*P>z*	*[95% Conf.*	*Interval]*
No. of Authors	0.096	1.101	0.025	3.92	0.000	0.048	0.144
No. of Countries	0.244	1.277	0.119	2.06	0.04	0.011	0.477
Funding	0.733	2.082	0.104	7.08	0.000	0.53	0.936
Readability of Abs.	-0.005	0.995	0.004	-1.24	0.214	-0.012	0.003
Constant	-0.684	0.505	0.187	-3.67	0.000	-1.049	-0.318
NB model	*Coef.*	*Exp(Coef.)*	*Std. Err.*	*z*	*P>z*	*[95% Conf.*	*Interval]*
No. of Authors	0.019	1.019	0.019	0.97	0.331	-0.019	0.056
No. of Countries	0.071	1.073	0.095	0.75	0.456	-0.115	0.257
Funding	0.353	1.423	0.104	3.38	0.001	0.148	0.557
Readability of Abs.	-0.018	0.983	0.003	-5.11	0.000	-0.024	-0.011
Constant	0.464	1.59	0.188	2.46	0.014	0.094	0.833
alpha	0.87	2.387	0.189	4.6	0.000	0.499	1.241

Social Sciences In Social Sciences (2000-2009), the number of countries is neither a significant determinant of citation counts nor zero citations ($p > 0.05$). With respect to the negative binomial model, the positive significant coefficients of the number of authors and the number of institutions indicate their association with increased citations. The expected mean citation count increased by 8.7% for each extra author and by 5.1% for each extra institution. With respect to the logit model, a one-unit change in the number of authors and the number of institutions decreases the mean zero citation by 12.8% and 11.3% respectively. Abstract readability associates with decreased citation counts, although it has no practical significance. Moreover, with respect to the logit model, there is no significant association between this variable and zero citations (Table 5).
In Social Sciences (2009), research funding was also taken into account. Results show that there is no significant association between research funding and citation counts, although this variable is associated with 80% decrease in zero citations. The results for the number of authors and the number of countries in the one-year model are similar to the results of the ten-year model but the abstract readability is behaving differently in the one-year model. This variable is significantly associated with increased citations in Social Sciences (2009) and a one-unit change in the abstract readability increased the citation counts by 0.5% (Table 6).

The overdispersion parameters are significant in all three models further justifying the negative binomial model (p for alpha<0.001).

Table 5. The results of hurdle model in Social Sciences (2000-2009)

Logit model	*Coef.*	*Exp(Coef.)*	*Std. Err.*	*z*	*P>z*	*[95% Conf. Interval]*	
No. of Authors	0.12	1.128	0.017	7.21	0.000	0.088	0.153
No. of Institutions	0.107	1.113	0.033	3.24	0.001	0.042	0.172
No. of Countries	0.024	1.024	0.066	0.36	0.717	-0.105	0.153
Readability of Abs.	-0.002	0.998	0.002	-1.25	0.212	-0.005	0.001
Constant	0.616	1.851	0.076	8.06	0.000	0.466	0.765
NB model	*Coef.*	*Exp(Coef.)*	*Std. Err.*	*z*	*P>z*	*[95% Conf. Interval]*	
No. of Authors	0.083	1.087	0.013	6.54	0.000	0.058	0.108
No. of Institutions	0.049	1.051	0.023	2.13	0.033	0.004	0.095
No. of Countries	0.023	1.023	0.044	0.52	0.603	-0.064	0.11
Readability of Abs.	-0.003	0.997	0.001	-2.09	0.037	-0.006	0.000
Constant	1.133	3.104	0.067	16.95	0.000	1.002	1.264
alpha	1.314	3.72	0.058	22.71	0.000	1.2	1.427

Table 6. The results of hurdle model including research funding in Social Sciences (2009)

Logit model	*Coef.*	*Exp(Coef.)*	*Std. Err.*	*z*	*P>z*	*[95% Conf. Interval]*	
No. of Authors	0.14	1.15	0.024	5.8	0.000	0.093	0.187
No. of Institutions	0.174	1.19	0.43	4.06	0.000	0.09	0.258
No. of Countries	0.241	1.272	0.101	2.37	0.018	0.041	0.441
Funding	0.589	1.802	0.137	4.28	0.000	0.319	0.859
Readability of Abs.	0.0005	1.0005	0.003	0.16	0.87	-0.006	0.007
Constant	-0.769	0.464	0.159	-4.85	0.000	-1.080	-0.458
NB model	*Coef.*	*Exp(Coef.)*	*Std. Err.*	*z*	*P>z*	*[95% Conf. Interval]*	
No. of Authors	0.048	1.049	0.013	3.53	0.000	0.021	0.074
No. of Institutions	0.057	1.058	0.026	2.21	0.027	0.006	0.109
No. of Countries	0.125	1.133	0.146	0.86	0.391	-0.161	0.412
Funding	0.074	1.076	0.199	0.37	0.71	-0.317	0.465
Readability of Abs.	0.005	1.005	0.002	1.99	0.046	0.00008	0.01
Constant	-3.423	0.033	5.684	-0.6	0.547	-14.563	7.718
alpha	3.28	26.57	2.75	1.19	0.049	-2.11	8.69

Discussion and conclusions

The analysis of the factors affecting citation counts of the papers that are cited at least once indicates that one component of research collaboration, the number of authors, is the only factor associated with increased citations in the ten-year model in the three subject fields. This factor also very significantly associates with decreased zero citations. Conversely, however, a study of a specific journal in Chemistry found no correlation between the number of authors and increased citation counts. This difference may result from the difference between the micro-level and macro-level analyses (Bornmann, Schier, Marx, & Daniel, 2012) or the smaller sample size for the single journal studied giving insufficient statistical

power to identify the association. The number of authors has not been found to be a significant determinant of citations in social and personality psychology (Haslam, et al., 2008). The authors believed that team-working is not necessarily a true reflection of research collaboration in this field.

Table 7. The results of extra hurdle models (only the negative binomial part) for the effect of the number of institutions on citation counts using a range of different fixed numbers of authors and countries (e.g., 3au_2cnty means 3 authors from 3 different countries)

Biology & Biochemistry					*Chemistry*				
Status	*Coef.*	*Exp (coef.)*	*P>|z|*	*Sample Size*	*Status*	*Coef.*	*Exp (coef.)*	*P>|z|*	*Sample Size*
2au_1cnty	-0.044	0.96	0.52	1935	2au_1cnty	-0.27	0.76	0.00	2562
3au_1cnty	-0.054	0.95	0.05	2307	3au_1cnty	-0.168	0.85	0.00	3090
4au_1cnty	-0.098	0.91	0.01	2144	4au_1cnty	-0.11	0.90	0.02	2686
5au_1cnty	-0.0003	0.99	0.9	1772	5au_1cnty	-0.065	0.94	0.18	1713
6au_1cnty	0.055	1.06	0.01	1315	6au_1cnty	-0.102	0.90	0.05	1008
7au_1cnty	-0.017	0.98	0.7	864	7au_1cnty	-0.1	0.90	0.14	505
8au_1cnty	-0.102	0.90	0.04	499	8au_1cnty	0.08	1.08	0.48	188
9au_1cnty	0.0054	1.01	0.9	325	9au_1cnty	0.03	1.03	0.74	135
10au_1cnty	0.125	1.13	0.1	199	10au_1cnty	0.028	1.03	0.8	67
3au_2cnty	0.02	1.02	0.85	377	3au_2cnty	0.03	1.03	0.84	424
4au_2cnty	-0.125	0.88	0.2	452	4au_2cnty	-0.069	0.93	0.53	513
5au_2cnty	0.02	1.02	0.68	448	5au_2cnty	-0.056	0.95	0.5	448
6au_2cnty	-0.11	0.90	0.04	423	6au_2cnty	-0.25	0.78	0.01	289

No clear evidence of the number of institutions was found in both Chemistry and Biology & Biochemistry but in Social Sciences, the number of institutions is a significant determinant of both increased citation counts and decreased zero citations in this field.

The number of countries has been a significant factor for increased citations in the majority of previous studies except for an institutionally-limited investigation of Harvard University. This university is one of the world's top universities and it seems logical in this context that its researchers benefit more from institutional collaboration than from international collaboration (Gazni & Didegah, 2010). However, the results of current study also show that the number of countries is a significant determinant of increased citation counts in none of the three fields.

The contradiction between the results of this study and some previous studies of international and institutional collaboration may result from the limited geographical and institutional coverage of previous research whereas the current study has a global coverage and seeks results at a macro-level. This study goes beyond a simple correlation between a predictor variable and citation counts. A co-analysis of predictors is considered here and the results are therefore more reliable although factors not considered in the analysis may also influence the results. Furthermore, the influence of research collaboration on research citation

impact is not uniform and varies across domains particularly for the institutional and international types of collaboration (Gazni & Didegah, 2010; Sooryamoorthy, 2009). However, the general tendency acknowledges the positive impact of the number of authors on the citation counts in all fields of science (Franceschet & Costanini, 2010).

Receiving grants from funding agencies brings many advantages to the research community by supporting researchers and paving the way for creative and high quality research especially in equipment-based research fields. The current study found that while there is a very strong significant impact of funding on citation counts in Biology & Biochemistry and Chemistry, there is no significant association between this variable and citation counts in Social Sciences. Therefore, it seems that Life Sciences and Physical Sciences are performing much better than Social Sciences in the area of receiving grants and publishing high impact research papers. This is probably resulted from the different natures of the subject fields that are experiment and equipment-based such as Chemistry and Biology & Biochemistry or theory-based such as Social Sciences. Receiving funds is so vital to an experiment-based research to provide the required equipment and conduct its experiments. So, the result of a funded research in such a research field is considerably different from a non-funded research and high impact results for funded research in experiment-based subject fields are expected due to less limitation in accessing resources and equipment.

Previous studies have also found that funding is not a significant determinant of citation counts in Psychology (Haslam et al., 2008) and Information Science (Croin & Shaw, 1999) which are in Social Sciences category. However, there are more studies that found funded research to be more highly cited than unfunded research in Life and Physical Sciences such as Medical Education, Biomedical research, Material Sciences, and Physics (Read et al., 2007; Zhao, 2010; Lewison & Dawson, 1998; Jowkar, Didegah, & Gazni, 2011).

Abstract readability was found to be statistically a significant determinant of both decreased citation counts and increased zero citations in Biology & Biochemistry and Chemistry but its influence was too small to be of practical value in the examined fields. This is whilst this variable is significantly associated with increased citations in Social Sciences meaning that the easier the abstract, the more number of citations in this field. However, previous research confirmed a negative association between the readability of abstract and citation impact of publications in the top institutions of the world (Gazni, 2011). As with previous studies of readability of the texts, all readability measures have a common limitation; they do not consider the characteristics of readers. The readers of scientific papers are experts in their own fields and have prior knowledge and interest in them. Hence an abstract graded as difficult based on its Flesch score may not be difficult for the scholars of the field (Gazni, 2011). On the other hand, scholars may scan the abstracts for keywords to find if the paper is relevant rather than reading the entire abstract.

In conclusion, this study attempted to identify effective determinants of research citation impact for scholars to help them to choose styles of high impact research. Team working was found to be a good determinant of citation impact in all three fields examined and the more number of authors, the higher the citation impact of the paper. Inter-institutional collaboration was a determinant of higher citation impact in the Social Sciences while international collaboration associates with citation counts in none of the fields. Moreover, research funding is very significantly associated with increased citations in Biology & Biochemistry and Chemistry and abstract readability contributes to increased citation counts in Social Sciences. The results do not encourage seeking international partners in the disciplines studied and writing more readable abstracts as neither seem to favour citation impact in Biology & Biochemistry and Chemistry.

Acknowledgements

This research is part of the FP7 EU-funded project ACUMEN on assessing Web indicators in research evaluation.

References

Bottle, R. S., Rennie, J. S., Russ, S. & Sardar, Z. (1983). Changes in the communication of chemical information I: some effects of growth. *Journal of Information Science,* 6, 103-108.

Bornmann, L. & Daniel, H. D. (2007). Multiple publication on a single research study: does it pay? The influence of number of research articles on total citation counts in biomedicine. *Journal of American Society for Information Science*, 58, 1100-1117.

Bornmann, L., Schier, H., Marx, W. & Daniel, H-D. (2012). What factors determine citation counts of publications in chemistry besides their quality? *Journal of Informetrics*, 6, 11-18.

Boyack, K. W. & Klavans, R. (2005). Predicting the Importance of Current Papers. *Proceedings of ISSI 2005,* 335–342. Edited by P. Ingwersen and B. Larsen. July 24-28, Stockholm, Sweden.

Cameron, C. A. & Trivedi, P. K. (2001). *Essentials of Count Data Regression, A companion to theoretical Econometrics*. Oxford: Blackwell.

Cronin, B. & Shaw, D. (1999). Citation, funding acknowledgement and author nationality in four information science journals. *Journal of Documentation*, 55, 402–408.

Franceschet, M. & Costantini, A. (2010). The effect of scholar collaboration on impact and quality of academic papers. *Journal of Informetrics*, 4, 540-553.

Gazni, A. (2011). Are the abstracts of high impact articles more readable? Investigating the evidence from top research institutions in the world. *Journal of Information Science*, 37 (3), 273-281.

Gazni, A. & Didegah, F. (2010). Investigating Different Types of Research Collaboration and Citation Impact: A Case Study of Harvard University's Publications. *Scientometrics*, 87(2), 251-265.

Gazni, A., Sugimoto, C. R., & Didegah, F. (2012). Mapping world scientific collaboration: authors, institutions, and countries. *Journal of the American Society for Information Science & Technology (JASIS&T),* 63(2), 323-335.
Glänzel, W. (2001). National characteristics in international scientific co-authorship relations. Scientometrics, 51(1), 69-115.
Glänzel, W. & Schubert, A. (2001). Double effort=Double impact? A critical view at international co-authorship in chemistry. *Scientometrics*, 50(2), 199-214.
Goldfinch, S., Dale, T., & DeRouen, K. (2003). Science from the periphery: Collaboration, networks and "periphery effects" in the citation of New Zealand Crown Research Institutes articles, 1995–2000. *Scientometrics*, 57(3), 321–337.
Hartley, J., & Benjamin, M. (1998). An evaluation of structured abstracts in journals published by the British Psychological Society. *British Journal of Educational Psychology*, 68(3), 443-456.
Hartley, J., Sotto, E., Pennebaker, J. (2002). Style and substance in psychology: Are influential articles more readable than less influential ones? *Social Studies of Science,* 32, 321-334.
Hartley, J. & Sydes, M. (1997). Are structured abstracts easier to read than traditional ones? *Journal of Research in Reading,* 20(2), 122-136.
Hartley, J. & Trueman, M. (1992). Some observations on using journal articles in the teaching of psychology. *Psychology Teaching Review,* 1(1), 46-51.
Haslam, N. et al. (2008). What makes an article influential? Predicting impact in social and personality psychology. *Scientometrics,* 76(1), 169-185.
Hilbe, J. M. (2011). *Negative Binomial Regression, second edition* (5th print, 2012). Cambridge, UK: Cambridge University Press.
Jowkar, A., Didegah, F. & Gazni, A. (2011). The Effect of Funding on Academic Research Impact: A case study of Iranian publications. *Aslib Proceedings*, 63(6): 593-602.
Katz, J. S., & Hicks, D. (1997). How much is a collaboration worth? A calibrated bibliometric model. *Scientometrics*, 40(3), 541-554.
Lariviere, V. & Gingras, Y. (2010). On the relationship between interdisciplinarity and scientific impact. *Journal of the American Society for Information Science and Technology*, 61, 126-131.
Leimu, R., & Koricheva, J. (2005a). What determines the citation frequency of ecological papers? *Trends in Ecology & Evolution*, 20, 28-32.
Leimu, R., & Koricheva, J. (2005b). Does scientific collaboration increase the impact of ecological articles? *BioScience*, 55, 438-443.
Levitt, J. M. (2011). Are funded articles more highly cited than unfunded articles? A preliminary investigation. *Proceedings of the ISSI 2011 conference*, Durban, South Africa, 1013-1015.
Lewison, G. & Dawson, G. (1998). The effect of funding on the outputs of biomedical research. *Scientometrics*, 41, 1-2, 17-27.

Lovaglia, M.J. (1989). Status characteristics of journal articles for editor's decisions and citations. *The Society for Social Studies of Science Annual Meeting*, November 15- 18, University of California at Irvine, Irvine, CA.
Milojević, S. & Leydesdorff, L. (in press). Information Metrics (iMetrics): A Research Specialty with a Socio-Cognitive Identity? *Scientometrics*.
Nagaoka, S., Igami, M., Eto, M., & Ijichi, T. (2011). *Knowledge Creation Process in Science: Basic findings from a large-scale survey of researchers in Japan*. Report by National Institute of Science and Technology Policy (NISTEP), Institute of Innovation Research (IIR), Hitotsubashi University.
Narin, F. & Whitlow, E. S. (1990). *Measurement of scientific co-operation and co-authorship in CEC related areas of science*. Report EUR 12900. Office for Official Publications in the European Communities, Luxembourg.
Narin, F., Stevens, K., & Whitlow, E. S. (1991). Scientific cooperation in Europe and the citation of multinationally authored papers. *Scientometrics*, 21(3), 313-323.
Persson, O., Glänzel, W., & Danell, R. (2004). Inflationary bibliometric values: The role of scientific collaboration and the need for relative indicators in evaluative studies. *Scientometrics*, 60(3), 421-432.
Peters, H. P. F. & van Raan, A. F. J. (1994). On Determinants of Citation Scores: A Case Study in Chemical Engineering. *Journal of the American Society for Information Science*, 45(1), 39-49.
Reed, D. A. et al. (2007). Association between Funding and Quality of Published Medical Education Research. *Journal of American Medical Association*. 298(9), 1002-1009.
Sooryamoorthy, R. (2009). Do types of collaboration change citation? Collaboration and citation patterns of South African science publications. *Scientometrics*, 81(1), 177–193.
Stremersch, S. Verniers, I. & Verhoef, P. C. (2007). The quest for citations: drivers of article impact. *Journal of Marketing,* 71, 171-193.
Thomson Reuters Technical Support (2013). Available at: http://ip-science.thomsonreuters.com/support/.
Zhao, D. Z. (2010). Characteristics and impact of grant-funded research: a case study of the library and information science field. *Scientometrics*, 84(2), 293-306.
Zimmerman, J. L. (1989). Improving a manuscript's readability and likelihood of publication. *Issues in Accounting Education*, 4, 458-466.

POSTERS

THE 2-YEAR MAXIMUM JOURNAL IMPACT FACTOR

María Isabel Dorta-González[1] and Pablo Dorta-González[2]

[1] *isadorta@ull.es*
Universidad de La Laguna (Spain)

[2] *pdorta@dmc.ulpgc.es*
Universidad de Las Palmas de Gran Canaria (Spain)

Introduction

During decades, the journal impact factor (JIF) has been an accepted indicator in ranking journals. The 2-year journal impact factor (2-JIF) counts citations to one and two year old articles, while the 5-year journal impact factor (5-JIF) counts citations from one to five year old articles. However, there are increasing arguments against the fairness of using the JIF as the sole ranking criteria (Althouse et al., 2009; Bensman, 2007; Bornmann & Daniel, 2008).

These indicators are not comparable among fields of science for two reasons: (i) each field has a different impact maturity time, and (ii) because of systematic differences in publication and citation behaviour across disciplines. Citation-based bibliometric indicators need to be normalized for such differences in order to allow for meaningful between-field comparisons of citation impact (Dorta-González & Dorta-González, 2010, 2011).

There is not an optimal fixed impact maturity time valid for all the fields. In some of them two years provides a good performance whereas in others three or more years are necessary. Therefore, there is a problem when comparing a journal from a field in which impact matures slowly with a journal from a field in which impact matures rapidly.

In this paper, we provide a source normalization approach based on variable citation time windows and we empirically compare this with the traditional normalization approach based on a fixed target window.

The variable citation time window

The delimitation among fields of science has until now remained an unsolved problem because these delineations are fuzzy at each moment in time and develop dynamically over time.

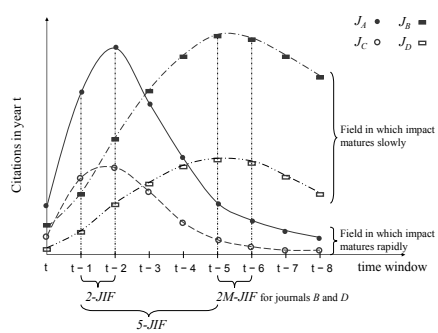


Figure 1. Citations distribution of journals

The choice for a variable rather than a fixed citation time window is based on the observation that in many fields citations have not yet peaked after 2 years, and in other fields citations have peaked long before 5 years. Therefore,

the application of a 2-year variable window is the optimal compromise for fields in which impact matures slowly in reaching its maximum citations while not penalising fields in which impact matures rapidly.

Figure 1 shows the citations distribution of four journals with different performance. Journals A and C belong to a field in which impact matures rapidly, while journals B and D belong to a field in which impact matures slowly. Then A has greater impact than C, and B has greater impact than D. Nevertheless, which journal has greater impact, A or B? And, C or D?

We define the rolling impact factors in year *t* of journal *i* as:

$$R_j\text{-}JIF_t^i = \frac{NCit_{t,\,t-j}^i + NCit_{t,\,t-j-1}^i}{NArt_{t-j}^i + NArt_{t-j-1}^i},\ j = 1,\ldots,h,$$

and the *2-year maximum journal impact factor* in year *t* of journal *i* as following:

$$2M\text{-}JIF_t^i = max_{j=1,\ldots,h}\left\{R_j\text{-}JIF_t^i\right\}.$$

The idea is to consider, for each journal, the citation time window with the highest average number of citations (i.e., the most advantageous period for each journal).

Materials and Methods

The bibliometric data was obtained from the online version of the 2011 Journal Citation Reports (JCR) Science edition. In the comparative analysis one journal category from each of the eight clusters obtained by Dorta-González & Dorta-González (2013) were considered. This was done in order to obtain journals with systematic differences in publication and citation behaviour. A total of 618 journals were considered. The categories and the number of journals are as follows: Astronomy & Astrophysics (56); Biology (85); Ecology (134); Engineering, Aerospace (27); History & Philosophy of Science (56); Mathematics, Interdisciplinary Applications (92); Medicine, Research & Experimental (112); and Multidisciplinary Sciences (56).

Table 1. Journal impact factors

Journal title	*2-JIF*	*2M-JIF*	*5-JIF*
AIAA J	1.057	1.458	1.277
AM NAT	4.725	5.750	5.280
ANN NY ACAD SCI	3.155	3.372	2.997
ASTRON ASTROPHYS	4.587	4.587	3.979
ASTROPHYS J	6.024	6.024	5.102
BIOL PHILOS	1.203	1.714	1.360
BIOMETRIKA	1.913	3.141	2.575
BRIT J PHILOS SCI	1.097	1.587	1.364
ECOLOGY	4.849	6.868	6.007
ECONOMETRICA	2.976	6.721	4.700
EXP HEMATOL	2.905	3.497	3.088
FASEB J	5.712	6.875	6.340
HIST SCI	0.667	0.818	0.699
IEEE T AERO ELEC SYS	1.095	2.288	1.680
J ECONOMETRICS	1.349	3.297	2.496
J GUID CONTROL DYNAM	0.941	1.370	1.159
LIFE SCI	2.527	2.880	2.732
P NATL ACAD SCI USA	9.681	11.167	10.472
P ROY SOC A-MATH PHY	1.971	2.086	1.987
PHYS REV D	4.558	4.558	4.027
PLOS ONE	4.092	5.756	4.537
STRUCT EQU MODELING	4.710	11.965	7.195
TRENDS ECOL EVOL	15.748	18.335	16.981
VACCINE	3.766	4.163	3.700

Results and discussion

Table 1 shows a sample of 24 randomly selected journals from those with the greatest overall impact (total citations) in eight JCR categories. Notice the ampleness in the interval of variation for each indicator. The 2-JIF varies from 0.667 to 15.748, while 2M-JIF varies from 0.818 to 18.335, for example. The general pattern is an increment in the 2M-JIF. However, this increment is in percentage terms much higher in the smaller values. This effect produces a concentration of data in the case of 2M-JIF, and consequently a reduction in the

variance. Table 2 shows the central-tendency measures for the aggregate data. It also shows the between-group variances. Note that all target windows reduce the between-group variance. However, the maximum citation time window produces the greatest reduction (3.203). Thus, this normalization by variable target windows reduces the between-group variance over 80%, when compared to within-group variance.

Table 2. Central-tendency and variability measures

Measures	*2-JIF*	*2M-JIF*	*5-JIF*
Median	1.245	1.745	1.531
Mean	2.142	2.827	2.481
Within-group variance	3.203	3.998	3.505
Between-group variance	0.709	0.795	0.728
Reduction in variance	2.494	3.203	2.777

Conclusions

Different scientific fields have different citation practices and citation-based indicators need to be normalized for such differences. In this paper, we provide a source normalization approach, based on a variable target window and we compare it with a traditional normalization approach based on a fixed target window.

The empirical application shows that our maximum citation time window reduces the between-group variance in relation to the within-group variance more than the rest of the indicators analyzed.

Finally, the journal categories considered are in very different areas in relation to the impact maturity time. Some of them are penalized by the 2-JIF and favored by the 5-JIF, and vice versa. This is the main reason why it is necessary to be cautious when comparing journal impact factors from different fields. In this sense, our index has behaved well in a great number of journals from very different fields.

References

Althouse, B. M., West, J. D., Bergstrom, C. T., & Bergstrom, T. (2009). Differences in impact factor across fields and over time. *Journal of the American Society for Information Science and Technology*, 60(1), 27–34.

Bensman, S. J. (2007). Garfield and the impact factor. *Annual Review of Information Science and Technology*, 41(1), 93–155.

Bornmann, L., & Daniel, H. D. (2008). What do citation counts measure? A review of studies on citing behavior. *Journal of Documentation*, 64(1), 45–80.

Dorta-González, P., & Dorta-González, M. I. (2010). Indicador bibliométrico basado en el índice h. *Revista Española de Documentación Científica*, 33(2), 225–245.

Dorta-González, P., & Dorta-González, M. I. (2011). Central indexes to the citation distribution: A complement to the h-index. *Scientometrics*, 88(3), 729–745.

Dorta-González, P., & Dorta-González, M. I. (2013). Comparing journals from different fields of science and social science through a JCR subject categories normalized impact factor. *Scientometrics* (in press). DOI 10.1007/s11192-012-0929-9

ACCURACY ASSESSMENT FOR BIBLIOGRAPHIC DATA

Marlies Olensky

marlies.olensky@ibi.hu-berlin.de
Humboldt-Universität zu Berlin, Berlin School of Library and Information Science, Unter den Linden 6, 10099 Berlin (Germany)

Introduction

While a variety of bibliometric and bibliographic studies compared the two main bibliometric data sources Web of Science (WoS) and Scopus (e.g. Meho & Yang, 2007) in terms of coverage, overlap, citation counts, etc., only a few studies (also) investigated the underlying problem: data quality of the data values (e.g. Hildebrandt & Larsen, 2008). What do we know about the data quality situation in WoS and Scopus? And what is the impact of data quality on citation analysis? In order to answer these questions, we need to take one step back and find a suitable method to assess data quality in these sources. This study investigates if an automated assessment, as described in the data quality literature, could be used to evaluate one aspect of data quality, namely data accuracy, for bibliographic data. The data accuracy of two bibliographic datasets from WoS and Scopus is assessed and compared to a manual assessment method. The results contribute to the research on determining the impact of data accuracy on citation analysis.

Inaccuracies in bibliometric data sources

Moed & Vriens (1989) conducted a study on the accuracy of citation counts, pitfalls during data collection and the influence of random and systematic errors on the citation analysis. They outline errors and variations occurring in the fields *author name*, *journal title*, *publication year*, *volume* and *starting page number*. In 2008, Hildebrandt & Larsen as well as Larsen, Hytteballe Ibanez & Bolling carried out two related studies on errors in the WoS. Larsen, Hytteballe Ibanez & Bolling (2008) investigated WoS' automatic matching and linking algorithm, identified patterns of errors and came up with improvements to the algorithm. The overall results showed that of 33,024 citations 6.2% were erroneous with at least one error. Hildebrandt & Larsen (2008) took a closer look at the high error rate in the field of Law. They found the most common errors in WoS were in the fields *cited page*, *author names* and *year*. On the whole, none of the previous studies looked further into finding a standardized method of how these errors could be rated and there is no distinct framework that allows a systematic analysis of data quality in bibliometric data sources.

Data quality / accuracy assessment

The literature provides a variety of techniques to assess data quality in databases and summarises those in different data quality assessment frameworks (e.g. Batini, Cabitza, Cappiello et al., 2008). However, before applying one or more of these frameworks to bibliographic data, one

should test if they are suitable for this purpose. Most of the frameworks follow Redman's definition of data quality (1996, pp. 245). This study chooses to investigate data accuracy as one of the four dimensions of data quality of data values (the others are completeness, consistency and timeliness).

Data quality (DQ) literature defines data accuracy basically as the ratio of correct and incorrect values. A slightly more complex way to calculate data accuracy is to measure the distance between the values stored in the database and those assumed to be correct. Among others, Redman (1996) suggests the Levenshtein or the Jaro-Winkler distance function. The Levenshtein score equals the number of single-character edits to turn one value into the other, whereas the Jaro-Winkler score measures the similarity of two strings. Ideally, the values from the database are measured against the real-world objects.

Research questions and data sample

The author addresses the following research questions:

- Do the data sources differ with regard to their data accuracy scores when assessed according to the DQ method and manually?
- As opposed to using the original publication for verification, could the publication list on the institutional website serve as silver standard to simplify the data accuracy assessment?

WoS and Scopus provide good coverage of chemistry literature that is why two Nobel Prize winners, one English-speaking and one from a non-English speaking country, from that domain were chosen: Roger D. Kornberg, an American biochemist and Professor at Stanford University School of Medicine and Gerhard Ertl, a German physicist and Professor emeritus at the Department of Physical Chemistry at Fritz-Haber-Institut der Max-Planck-Gesellschaft in Berlin, Germany. Publications of both Nobel Prize winners were retrieved for a publication period of 10 years (1998-2007), regardless whether they were the first or co-author. Articles and proceedings / conference papers were analysed, all other publication types were excluded.

Methodology

The full bibliographic records from WoS and Scopus were downloaded, investigated and compared to the original publication (*gold standard*) and the data from the institutional websites (*silver standard*). Bibliographic data accuracy is characterized by the data fields of *author name*(s) (including *first* and *second initial* of the given names), *article title*, *journal title*, *volume number*, *publication year* and *pagination* (Olensky, 2012), which are therefore used as assessment parameters. The automated evaluation calculated the Levenshtein distance score for each bibliographic field, comparing it to the original publication and the data from the institutional websites. The findings of a previous literature study (Olensky, 2012), where the author investigated the definition of what an inaccuracy is as well as the weighting of these in the different bibliographic fields, were considered during the manual assessment.The author analysed the discrepancies the automatic evaluation had found and assigned *inaccuracy points* (IAP) to the respective data fields in order to weigh inaccuracies according to their impact: 0 = accurate, 1 = minor, 2 = medium, 3 = major inaccuracy.For each assessment method, the scores were accumulated for each record per dataset and the total accuracy scores

were calculated. Additionally, the data source providing the lowest score per record was determined and accumulated per dataset.

Results and Conclusion

The main result is that the Levenshtein distance function is a good means to determine whether a data record contains discrepancies, but the score does not provide a true picture of how inaccurate a field is without the application of additional rules. Table 1 illustrates the accuracy scores from the manual and the automated evaluation.

Table 1. Accuracy scores compared to original publication and website, automated/manual assessment method, on record level.

	WoS		Scopus	
	Ertl	Kornb.	Ertl	Kornb.
Orig.aut.	30%	27%	37%	38%
Webs.aut.	38%	15%	8%	5%
Δ	8%	12%	29%	33%
Orig.man.	59%	73%	75%	90%
Webs.man.	55%	81%	59%	87%
Δ	4%	8%	16%	3%

The rules found in the course of the manual assessment reflect most of the required adjustments to be made to the automatic assessment method. Regarding the question whether the publication list from the institutional website could be used as *silver standard* instead of manually gathering the bibliographic data from the original publications, further analysis is needed and more websites need to be examined for their suitability.

In contrast to the accuracy scores on record level, the accuracy scores per field are high (≥ 95%). This proves that both, WoS and Scopus, provide very accurate data values and the difference between them is marginal. Both assessment methods confirmed that Scopus provides the most accurate data for both data samples. In future work, the author will apply the modified assessment method to a larger, more representative data sample that includes cited and citing publications. Also, the impact on citation analysis will be investigated with the ultimate goal of finding a standardized assessment of bibliographic data accuracy.

References

Batini, C., Cabitza, F., Cappiello, C., & Francalanci, C. (2008). A comprehensive data quality methodology for Web and structured data. *International Journal of Innovative Computing and Applications, 1*(3), 205–218.

Hildebrandt A.L., Larsen B. (2008). Reference and Citation Errors – a Study of Three Law Journals.Presentation of student work at the Royal School of Library and Information Science. Copenhagen, Denmark.

Larsen B., Hytteballe Ibanez K., Bolling P. (2008). Error Rates and Error Types for the Web of Science Algorithm for Automatic Identification of Citations. Presentation of student work at the Royal School of Library and Information Science.Copenhagen, Denmark.

Meho, L. I., & Yang, K. (2007). Impact of Data Sources on Citation Counts and Rankings of LIS Faculty: Web of Science vs. Scopus and Google Scholar. *Journal of the American Society for Information Science and Technology, 58*(13), 2105–2125.

Moed, H. F., & Vriens, M. (1989). Possible inaccuracies occurring in citation analysis. *Journal of Information Science, 15*(2), 95-107.

Olensky, M. (2012). How is Bibliographic Data Accuracy Assessed? In É. Archambault, Y. Gingras, & V. Larivière (Eds.), *Proceedings of 17th International Conference on Science and Technology Indicators* (pp.628–639). Montréal.

Redman, T. C. (1996). Data quality for the information age. The Artech House computer science library. Boston Mass.: Artech House.

ANALYSIS OF SEARCH RESULTS FOR THE CLARIFICATION AND IDENTIFICATION OF TECHNOLOGY EMERGENCE (AR-CITE)

Robert K. Abercrombie, Bob G. Schlicher, and Frederick T. Sheldon

abercrombier@ornl.gov, schlicherbg@ornl.gov, and sheldonft@ornl.gov
Oak Ridge National Laboratory, 1 Bethel Valley Road, Oak Ridge, TN 37831-6085 (USA)

Introduction

This research examines emerging technologies from initial discovery (via original scientific and conference literature), through critical discoveries (via original scientific, conference literature and patents), transitioning through Technology Readiness Levels (TRLs) and ultimately on to commercial application. The purpose of this study is to address the relationships among multiple disparate sources of information as a way to explain systematically the emergence of new technologies from innovation on through to commercial application with regards to TRLs. In one example, we investigate the combinations of four distinct and separate searchable on-line networked sources (i.e., scholarly publications and citation, patents, news archives, and on-line mapping networks) as they are assembled to become one collective network (a data set for analysis of causal relations). In another example, we investigate the combinations of five categories of data sources (i.e., university R&D, industry R&D, product emergence, and two levels of annual market revenue [$1B (USD) and $10B (USD)]). These established networks and relationships form the basis analyze the temporal flow of activity (searchable events) for the multiple example subject domains that we investigated.

Background Related Work

The logical sequence of milestones is derived from our analysis of a previously documented data set and technology that includes the initial discovery (evident via original scientific and conference literature), the subsequent critical discoveries (evident via original scientific, conference literature and patents), and the transitioning through the various TRLs ultimately to commercial application (Abercrombie, Udoeyop, & Schlicher, 2012).

The TRL is defined as a measure used to assess the maturity of evolving technologies (devices, materials, components, software, work processes, etc.) during their development and in some cases during early operations ("Defense Acquisition Guidebook," 2012). TRLs can serve as a helpful knowledge-based standard and shorthand for evaluating and classifying technology maturity, but they must be supplemented with expert professional judgment.

Research Hypothesis and Experimental Design

The question we investigate is the following. “Can one map the life cycle of a technology in a standard

methodological way to quickly identify and classify the emergence of a specific technology?"

Using the technology evolution model (TEM) developed in our prior work (Abercrombie, et al., 2012), a step wise analysis was conducted, applying the definitions of the TRLs to the corresponding milestones in the TEM to two case studies. The first case study applied the TEM to investigate the original Simple Network Management Protocol (SNMP) data set spanning the years 1988–2008. The second case study conducted a similar step-wise TRL analysis on a data set, spanning the years 1965–2011 (Lee et al., 2012). This second study addressed the evolution of eight specific subject areas of fundamental research in Information Technology (IT) (Digital Communications, Computer Architecture, Software Technologies, Networking, Parallel & Distributed Systems, Databases, Computer Graphics, and AI & Robotics), and respective industry interest categories (Broadband & Mobile, Micro-processors, Personal Computing, Internet & Web, Cloud Computing, Enterprise Systems, Entertainment & Design, and Robotics & Assistive Technologies) documenting university research (equivalent to scholarly pursue) and industry R&D (represented in patents and trade secrets) evolution to products with their respective market share.

Results

The first data set consists of four distinct and separate searchable on-line networked sources (i.e., scholarly publications and citation, patents, news archives, and on-line mapping networks). The data reflects a time line of the technology transitions categorized by TRLs as shown in Figure 1. These results convinced us to adapt the TEM so that the TRLs are used to determine the stage in the sequence of evolution (technology transfer).

The second case study uses the modified TEM to study a data set created from survey data from the IT Sector (Lee, et al., 2012) measuring the relationships among universities, industry, and governments' innovations and leadership. Figure 2 identifies the TRL transitions on eight subject areas of fundamental research in IT and industry interest categories.

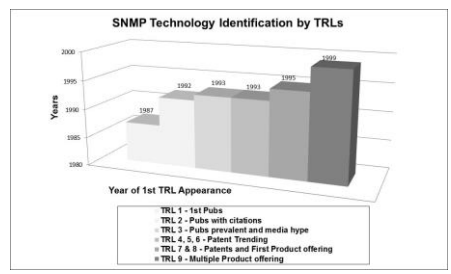


Figure 1. SNMP Technology Identification by TRLs.

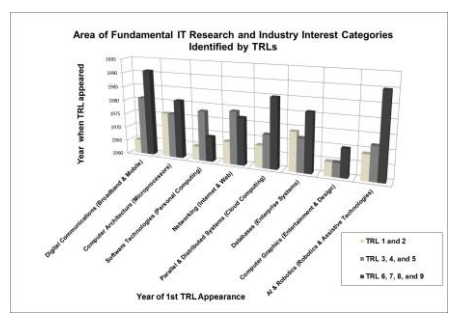


Figure 2. Area of Fundamental IT Research Technology Identification by TRLs.

Discussion and Conclusion

In general, when a new technology is first invented or conceptualized, it is not suitable for immediate application. Instead, new technologies are usually subjected to experimentation, refinement, and increasingly realistic

testing. However, applying the definitions of the TRL transitions provides a stepwise concrete explanation for the subject technology's evolution thus giving insight into its maturity and market impact.

This work examines and clarifies how to systematically classify the technology evolution starting from an initial discovery (via original scientific and conference literature), through critical developments (via original scientific, conference literature and patents), transitioning through TRLs to commercial application and significant economic impact. The relationships among multiple disparate sources of information were addressed, as a way to explain systematically the identifiable states of new technologies, from innovation on through to commercial application. In the first case study we selected a very well-known documented technology to test the TRL transitioning hypothesis. In the second case study the TRL transitioning hypothesis was validated by selecting a cross-section of fundamental IT research domains and their corresponding industry interest category spanning 1965–2011. In both studies, applying the TRL transitioning technique to the documented subject areas resulted in trends that clarified and refined the identification of the emergent technology. The TRL transitions, in the modified TEM, are useful in the creation of business intelligence. Business intelligence assists in providing a basis for strategic business decision(s). Further research is needed to refine the critical underlying data sources. In this study only two economic impact categories were used. To better under the progression (i.e., enabling breakthroughs) of technology transfer, it will be necessary to subdivide the economic impact categories into smaller bin sizes. Another area of investigation is to address the associated supply chains among the industry interest categories. We would like to better understand how inventions from one category affect (overlap) the emergence (development) properties in another category (e.g., phone hardware versus iTunes software). Moreover, we intend to investigate alternative techniques to better understand key agents of change (i.e., TRL transitioning) toward a more robust identification of technology emergence.

Acknowledgments

This manuscript has been authored by a contractor of the U.S. Government (USG) under contract DE-AC05-00OR22725. Accordingly, the USG retains a nonexclusive, royalty-free license to publish or reproduce the published form of this contribution, or allow others to do so, for USG purposes.

References

Abercrombie, Robert K., Udoeyop, Akaninyene W., & Schlicher, Bob G. (2012). A study of scientometric methods to identify emerging technologies via modeling of milestones. *Scientometrics, 91*(2), 327-342. doi: 10.1007/s11192-011-0614-4

Defense Acquisition Guidebook. (2012). Retrieved January 15, 2013, from https://acc.dau.mil/CommunityBrowser.aspx?id=518692&lang=en-US

Lee, Peter, Dean, Mark E., Estrin, Deborah L., Kajiya, James T., Raghavan, Prabhakar, & Viterbi, Andrew J. (2012). *Continuing Innovation in Information Technology*. Washington, DC: The National Academies Press.

APPLICATIONS AND RESEARCHES OF GIS TECHNOLOGIES IN BIBLIOMETRICS

Wang Xuemei[1, 2], Ma Mingguo[2], Li Xin[2], Zhang Zhiqiang[1]and Ma Jianxia[1]

[1]*wxm@lzb.ac.cn; zhangzq@lzb.ac.cn; majx@lzb.ac.cn*
Lanzhou Branch of the National Science Library, Scientific Information Center for Resources and Environment, Chinese Academy of Sciences, Lanzhou, 730000 (China)

[2]*mmg@lzb.ac.cn; lixin@lzb.ac.cn*
Cold and Arid Regions Environmental and Engineering Research Institute, Chinese Academy of Sciences, Lanzhou, 730000 (China)

Introduction

The literatures normally involve some spatial related information. The basic information source is the authors' institutes, cities and countries of articles, citing and cited articles. For example, the study areas, sampling points, observing points, sampling units, sampling strips are this kind of information.

Geographic Information System (GIS) integrates hardware, software, and data for capturing, managing, analyzing, and displaying all forms of geographically referenced information. It is widely use in practically every profession. There are also some practical applications and initial research works in recent years. In this paper, the advance of applications and researches of GIS technologies in bibliometrics was reviewed. Then the potential and further works was discussed in the prospect part.

Spatial information mining from the literature

For the spatial analysis and display, the raw spatial information is the important base and need to be extracted from large mass of literature. Some information with the regular structure can be extracted automatically and quickly. For example, the author affiliations, cities and countries of author and co-authors are normally included in the literature. Because the names of cities and states have fixed formats, they can be linked with the vector data within the spatial distribution information (Wang & Ma, 2009). The author affiliations are also relatively customary. Sometimes the shorter form or subsidiaries are used, which can be normalized based on the unit dictionary.

There are abundant toponym (place name) information in the abstracts and texts of the literature. These toponym data can also be linked with the vector data. But the toponym data is very complicated. There are various place forms, such as towns, mountains, rivers, lakes, or streets. But the least of the difficulties is the linkage between the place names with the map coordinates reference place points or regions. Therefore the researches can focus on some thematic place names.

Spatial display and basic spatial operations

Spatial display and spatial query are the basic operations of the GIS tools. The GIS is used directly in the digital library construction. With the development of the WEBGIS, more and more literature

database websites begin to use WEBGIS to display the spatial distribution of the authors, the spatial relationship among the co-authors and citations. For example, the BioMedExperts, AuthorMapper were developed byGoogle Earth or Google map. Some other browsing tools were developed for geographic digital library, such as GeoVIBE andLitmap.

The literature citation of a research unit (e.g. researcher, laboratory, institute, city or country) can also be displayed on spatial map. The interannual changes can be shown by using histogram or trend line for each spatial vector feature. When the spatial feature was queried, the linked database of multi-year data can be shown. The patent is also an important literature format. The USPTO's patent data are mapped and displayed using overlays to Google Maps. The overlays indicate both the quantity and quality of patents at the city level (Leydesdorff&Bornmann, 2012).

On the contrary, it is difficult to display spatial-linked information with the irregular structure on the internet. The most important step is the information extraction of other items with the irregular structure, which is the work focus in the future. After this step, the spatial information linked with the vector data was extracted from the literature, it could be inputted into the GIS tools.

Spatial analysis

Spatial analysis is the process to conduct spatial data various kinds of handling, and then get information, clues and knowledge from that. Here it can advance the deep data mining of the spatial-linked information from the literature. GIS provides the specialized platforms for computing spatial relationships among spatial units. The most commonly used spatial analysis methods include tracking analysis, buffer analysis, overlay analysis, route analysis, network analysis, spatial interpolation, geostatistical analysis. The application of the GIS spatial analysis in Bibliometrics is still in its beginning stage.

For example, the spatial interpolation method can be used to find the high density regions based on the highly-cited paper numbers. It can also be used to detect the "hot regions" of the scientific publication and citation (Bornmann&Waltman, 2011). The geostatistical analysis was used in the geostatistics literature indexed from Web of Science, Scopus and Google Scholar (Heng, Minasny& Gould, 2009). The buffer regions were used to analyze the spatial distribution of the sampling points in Qinghai-Tibet Plateau as the increasing distance with the traffic lines, such as railway and highway (Wang, Ma, Li, & Zhang, 2012).

Quantitative indexes

Some Bibliometrics indexes were developed to quantify the characteristics and amounts of the literatures.When these quantitative indexes of the research units are linked with the spatial positions of these research units, they can be displayed, queried and retrieved spatially. Some newly developed quantitative indexes with geographical position and direction are more suitable for spatial representation and analysis by using GIS technologies.The distance factor is also used to measure the spatial distribution pattern of the bidirectional knowledge flow. The spatial distance is calculated among the citing or cited papers based on the GoogleMapAPI and Yahoo PlaceFinder. But the relative researches are still in the initial stage and made in recent years.

Conclusions and discussion

For promoting the application and research of the GIS in Bibliometrics, there are some works can be carried out in the future.

(1) Application of regular information structure. The regular information structure can be queried and calculated instantly and automatically. If there is more geographically relative information with regular structure in the articles, it is easier for literature database to display its resources using the GIS platforms. Some international standards can be set up especially for the scientific publication. The authors and their affiliation information need be standardized in formats and word use. The spatial position involved in main text can also be standardized with the unified templates. But there are differences of the written requirement among various journals.

(2) Preparation of basic thematic maps. For the researchers who are engaged in the Bibliometrics, it is difficult to collect and arrange the basic thematic maps for linking the geographically relative information. Even though there are some WEBGIS resources can be used openly, some special applications need more preparation work. For a given study area (e.g. Qinghai-Tibet Plateau, Amazon Basin), there are a lot of hot spot position names to call the local mountains, lakes, rivers, villages, and so on. These position names are normally not included in the general map platforms, which need more collection and arrangement works.

(3) Development of information recognition tools. It is a key process to extract the geographically relative information from numerous scientific articles for using GIS in the Bibliometrics. There are seldom special tools to realize information recognition easily. Some tools can be developed currently for the thematic information recognition. It is also need to develop a tool to extract the coordinate information with latitude and longitude position and output to the standard geographic format.

(4) Development of the internet display platforms. There are a lot of internet electric maps online, which are widely used and developed fast. We can use these platforms to expand the functions to afford the information display and query of spatial literature information. More functions can be released especially for the application of GIS in literature representation. If possible, a specifically designed internet display platform is excepted to developed for realizing this application needs.

Acknowledgments

This work was supported by Special Project of Literature Ability of Chinese Academy of Sciences, and the National Natural Science Foundation of China (grant no.: 40701133).

References

Bornmann L. &Waltman L. The detection of “hot regions” in the geography of science—A visualization approach by using density maps. Journal of Informetrics 5 (2011c) 547– 553.

Heng, T., Minasny, B. & Gould, M. (2009).A geostatistical analysis of geostatistics.Scientometrics, 80(2): 491-514.

Leydesdorff L. &Bornmann L. Mapping (USPTO) Patent Data using Overlays to Google Maps. Journal of the American Society for Information Science and Technology 63(7) (2012) 1442-1458.

Wang, X. M. & Ma, M. G. (2009).Spatial information mining and visualization for Qinghai-Tibet Plateau's literature based on GIS.

Proceedings of International Symposium on Spatial Analysis, Spatial-Temporal Data Modeling, and Data Mining (ISSA 2009) (pp. 74920T, 1-8). Wuhan.

Wang, X. M., Li, X., Ma, M. G., & Zhang, Z. Q. (2012).Spatial analysis on the geographical information of the scientific literature for qinghai-tibet plateau. Advances in Earth Science, 27(11): 1288-1294.

APPROPRIATE COVERAGE OF SCHOLARLY PUBLISHING IN THE SOCIAL SCIENCES AND HUMANITIES - A EUROPEAN OVERVIEW

Gunnar Sivertsen[1], Elea Giménez-Toledo,[2] and Tim C. E. Engels[3]

[1] *gunnar.sivertsen@nifu.no*
Nordic Institute for Studies in Innovation, Research and Education (NIFU), Wergelandsveien 7, NO-0167 Oslo, Norway

[2] *elea.gimenez@cchs.csic.e*
G.I.de Evaluación de Publicaciones Científicas (EPUC), Centro de Ciencias Humanas y Sociales (CCHS) Spanish National Research Council (CSIC) C/ Albasanz, 26-28., 28037 Madrid, Spain

[3] *tim.engels@ua.ac.be*
Centre for R&D Monitoring (ECOOM), University of Antwerp, Middelheimlaan 1, B-2020 Antwerp, Belgium, and Antwerp Maritime Academy, Noordkasteel Oost 6, B-2030 Antwerp, Belgium

Introduction

Achieving full coverage of the scholarly publications in the social sciences and humanities (SSH) in bibliographic data sources is notoriously difficult (Hicks, 1999; Archambault et al., 2006; Nederhof, 2006). Although commercial databases such as the Web of Science (WoS) and Scopus have made considerable advances in increasing the coverage of the archival journals and articles in these fields, they still give limited representation of the SSH (Hicks and Wang, 2009) especially of output by researchers in non-English-speaking countries (Lariviére and Macaluso, 2011). In Flanders, Norway and Spain, however, attempts have been made to cover the scholarly output in the SSH and its publication channels more systematically and comprehensively (Sivertsen, 2010; Engels et al., 2012; Gimenez-Toledo et al., 2013).

In this poster, we will present an overview of how European countries manage to cover the publications in the social sciences and humanities in bibliographic databases for statistics, assessment and/or funding of research. It is our hypothesis that there are recent achievements with regard to better coverage in several countries. Still, they are not yet visible on the European level. Hence there is a need for an overview and maybe also a potential for collaboration between the initiatives.

Methods

We have performed a survey on email since February 2013 by contacting 28 colleagues in 28 countries and asking them to answer the questions cited below. If no response, we will contact other colleagues in the same country. So far, we have responses from 19 countries. We expect this number to increase before we finalize the poster, which we will design with the following elements:

- A map of Europe visualizing the main results from each country

- Tables summarizing in more detail the results from the survey
- A short text discussing the results and their implications.

The questionnaire

The questions covered by our survey are cited below:

1. In your country, are comprehensive bibliographic data (exceeding the journal coverage in ISI Web of Knowledge or Scopus) for scholarly publishing in the social sciences and humanities collected systematically (not only as individual publication lists) and continuously (i.e. not only as part of a survey) for the purpose of research information, statistics, assessment or funding?

   If no, you may reply without continuing to the other questions. If yes,

2. Are these publication types covered?
   a. Articles in peer-reviewed international journals
   b. Articles in peer-reviewed national journals
   c. Articles in edited scholarly books and book series
   d. Scholarly monographs
   e. Publications for students and non-academic audiences
3. For which purpose?
   a. Research information
   b. Statistics and studies
   c. Research assessment
   d. Project funding
   e. Institutional funding
   f. Full text repositories
   g. Other:

4. At which level are the data collected (please provide URL to relevant organization(s)?
   a. At the national level:
   b. At the institutional level:
   c. Other:
5. Are the researchers themselves providing and/or correcting their data?

6. Are the data complete from the point of view of the individual researcher?

7. Are the data available in a database that can be searched and analysed?
   a. If yes, supply URL:

8. If there are examples of published studies based on these data, please give one or more references:

9. Please name relevant organizations or persons that you would like us to know of:

10. Please add more information if you would like to do so (e.g. more details about the types of data).

Results and discussion

The preliminary result is that there are variations within Europe from countries that have achieved complete representations of the scholarly output in the SSH to countries with no representation at all. We observe that several countries in Eastern Europe have implemented database or current research information systems (CRIS) that cover SSH output to a large extent. In the Nordic countries too, several initiatives have successfully implemented. In large countries such as Germany, a more mixed pattern emerges, often with databases that are focused on one or more disciplines rather than the whole of the SSH. We conclude that in several countries clear progress has been made in achieving comprehensive coverage of SSH output. The full results of our survey will be presented and discussed in the poster.

References

Archambault, E. et al. (2006) 'Benchmarking Scientific Output in the Social Sciences and Humanities: The Limits of Existing Databases', *Scientometrics*, 68: 329–42.

Engels, T. C. E., Ossenblok, T. L. B. and Spruyt, E. H. J.(2012) 'Changing Publication Patterns in the Social Sciences and Humanities, 2000-2009', *Scientometrics*, 93 (2). 373-390.

Gimenez-Toledo, E., Tejada-Artigas, C., and Manana-Rodriguez, J. (2013). Evaluation of scientific books' publishers in social sciences and humanities: Results of a survey. Research Evaluation, 22 (1): 64-77.

Hicks, D. (1999) 'The Difficulty of Achieving Full Coverage of International Social Science Literature and the Bibliometric Consequences', *Scientometrics*, 44: 193–215.

Hicks, D. and Wang, J. (2009) Towards a Bibliometric Database for the Social Sciences and Humanities. A European Scoping Project (Appendix 1 to Martin et al., 2010). Arizona: School of Public Policy, Georgia University of Technology.

Lariviére, V. and Macaluso, B. (2011) 'Improving the Coverage of Social Science and Humanities Researchers' Output: The Case of the Érudit Journal Platform', *Journal of the American Society for Information Science & Technology*, 62: 2437–42.

Nederhof, A. J. (2006) 'Bibliometric Monitoring of Research Performance in the Social Sciences and the Humanities: A Review', *Scientometrics*, 66: 81–100.

Sivertsen, G. (2010). A performance indicator based on complete data for the scientific publication output at research institutions. *ISSI Newsletter*, 19 (6), 22-28.

ARE REGISTERED AUTHORS MORE PRODUCTIVE?

Sarah Heeffer[1], Bart Thijs[2], and Wolfgang Glänzel[3]

[1] *sarah.heeffer@kuleuven.be*
KU Leuven, ECOOM and Dept. MSI, Leuven (Belgium)

[2] *bart.thijs@kuleuven.be*
KU Leuven, ECOOM and Dept. MSI, Leuven (Belgium)

[3] *wolfgang.glanzel@kuleuven.be*
KU Leuven, ECOOM and Dept. MSI, Leuven (Belgium)
Library of the Hungarian Academy of Sciences, Dept. Science Policy & Scientometrics, Budapest (Hungary)

Introduction

The identification of authors in bibliographic databases and their assignment to research universities, research institutions or companies is still one of the big challenges in Scientometrics at the micro and meso level. Correct author identification is indispensable, above all, in longitudinal studies on scientific careers, studies of researchers' mobility or in monitoring constitution and performance of research teams (Strotman & Zhao, 2012). Recently the large abstract and citation databases Web of Science (Thomson Reuters) and Scopus (Elsevier) have introduced their ResearcherID and Author ID, respectively. Both are supposed to uniquely identify scientific authors but experience has taught us that these IDs are not yet fully implemented and that errors and multiple assignments are not quite the exception to the rule.

The present study aims at a systematic analysis of the cleanness of ResearcherIDs, their acceptance by authors and their implementation in the mirror of national research output and subject-specific peculiarities as reflected by major science fields. Finally we have analysed in how far ResearcherIDs can be used to represent national and field-specific publication-activity patterns. The latter question is important to find reference standards for publication activity such as otherwise only known for citation indicators so far.

Data sources and data processing

In order to use a reasonable publication set we have selected seven countries from Europe and one country from Asia. These countries are Austria, Belgium, Germany, Hungary, Netherlands, China, Switzerland and UK. All 'citable' documents with at least one author from these countries and one or more authors with ResearcherID (RID) have been downloaded from the 2009–2011 volumes of the online version of Thomson Reuters' (TR) *Web of Science* (WoS). It should be stressed that the author with RID needs not necessarily be affiliated with an institution in the countries in question. After download these papers have been matched with all publications from these countries extracted from the WoS custom-data set

licensed at ECOOM. In a following step all RIDs have been uniquely assigned to countries on the basis of TR's affiliation tag. RID's from foreign countries have been removed from the national sets. All authors without RID have also been assigned to countries and – as far as possible – disambiguated on the basis of name and first initial and affiliation. After the cleaning process a certain amount of homonyms and synonyms still remains in the data set as well as some uncertainty about the authors' consequent and correct mention of their identifiers. All papers have been assigned to major fields on the basis of the Leuven-Budapest classification scheme. Papers can be assigned to more than one field or country due to journal assignment and co-authorship, respectively.

Table 1. Shares of RID authors and papers with RID authors per country [Data sourced from Thomson Reuters Web of Knowledge]

Country	Papers	A (%)	B (%)	C (%)
AUT	36272	45.7	12.1	27.1
BEL	53682	42.8	13.7	28.4
DEU	277524	41.2	15.4	22.9
HUN	17073	49.6	20.9	31.6
NLD	97625	45.1	19.2	30.2
CHN	423510	36.5	13.0	26.4
CHE	69958	47.8	16.5	19.6
GBR	298857	48.8	12.5	27.7

Legend: A = Mean share of RID per paper, B = share of papers with RID, C = share of authors with RID

Methods and results

Researcher names associated with RIDs were matched with author names as they appear on the paper. This allowed us to identify some problems. First, RIDs are not only used by authors. Some institutes and author groups mark their publications by an RID. Some RIDs claim several papers while the researcher name does not match any of the authors. A RID is not always unique. Some authors have created and are using different RIDs to claim the same papers with these different RIDs. The overwhelming share of RIDs, however, seems to be created by individuals and used in a correct manner.

Table 1 displays the mean shares of authors with RID (A) and the share of papers (B) respectively authors (C) with an RID. On an average, 40%–50% of authors on a paper have a RID registration. In China we have found the lowest share, while Hungary and the UK have the highest one around 50%. National shares of papers with RID authors is much lower; it ranges between 12% and 21%. Here Hungary and the Netherlands are at the high end and the UK has jointly with Austria the lowest share. Similarly, Hungary and the Netherlands have the highest shares of registered authors but unlike the previous static, Germany and Switzerland form the low end here. Roughly one quarter to one third of all authors from the country selection use a RID registration. These effects are not the result of foreign collaboration since co-authors from other countries have been removed from the statistics.

Table 2. Mean publication activity of RID authors vs. authors in RID papers and all authors per country [Data sourced from Thomson Reuters Web of Knowledge]

Country	A	B	C
AUT	3.89	3.35	7.73
BEL	4.52	3.16	7.02
DEU	4.95	3.84	7.56
HUN	4.00	2.99	4.76
NLD	4.00	3.02	7.77
CHN	23.34	9.26	8.33
CHE	4.32	4.60	6.81
GBR	4.09	3.13	5.39

Legend: A = Mean activity of all authors, B = Mean activity of authors in RID papers, C = Mean activity of RID authors

The comparison of publication activity reveals other aspects of national patterns of RID use. The mean activity is certainly distorted by insufficient name disambiguation. Although the national statistics for all authors reflect similar activity for most countries (ranging from 4 to 5), China's extreme average activity points to identification issues.

Table 3. Mean publication activity of all authors (A) vs. RID authors (C) per major field [Data sourced from Thomson Reuters Web of Knowledge]

Field	A	C	Field	A	C
A	2.17	3.05	M	3.11	4.40
B	2.40	3.01	N	2.51	3.84
C	3.12	4.76	O	1.74	2.25
E	2.14	2.37	P	4.96	4.85
G	3.58	4.10	R	1.97	2.28
H	1.89	1.90	S	1.67	2.15
I	2.98	3.53	Z	2.48	3.53

Legend: A: agriculture & environment; B: biosciences (general, cellular & subcellular biology; genetics); C: chemistry; E: engineering; G: geosciences & space sciences; H: mathematics, I: clinical and experimental medicine I (general & internal medicine); M: clinical and experimental medicine II (non-internal medicine specialties); N: neuroscience & behavior; O: social sciences II (economical & political issues), P: physics; R: biomedical research; S: social sciences I (general, regional & community issues), Z: biology (organismic & supraorganismic level)

The mean activity of all authors in the RID set is generally somewhat lower but still in line with the activity of all authors. Here the Chinese value is more realistic. As expected, the activity of authors using RID (cf. column C in Table 2) is distinctly higher than the activity of all authors (except for China). However, China has still the highest activity, followed by the Netherlands, Austria and Germany. Of course, these values can be influenced by national publication profiles, therefor we have a look at subject-specific peculiarities of activity patterns before we have a closer look at the distribution of papers over authors using or not using RID. Because of the bias in the Chinese data, we have removed China in the following. Table 3 shows the mean activity (all authors vs. RID) for 12 major fields in the sciences and two fields in the social sciences. Again, the mean publication activity of RID authors generally exceeds that of the reference standard based on all authors. Physics forms the only exception. Also subject-specific peculiarities can be observed: mathematics and the social sciences have the lowest standards, followed by biomedical research and engineering. The deviation of the values presented in Table 3 from those in Table 2 are caused by the 'multidisciplinarity' of authors: RID authors are active in 2.5 fields on an average, while all authors in about 2.2 fields.

The mean activity of all authors in all fields combined amounts to 4.71, that of RID authors 6.87. Similarly, the corresponding share of authors with one paper amounts to 43.1% and 21.7%, respectively. Furthermore, RID authors are more productive at the high end of the distribution. The distribution is plotted in Figure 1. It goes without saying that the two distributions are distinctly different and it needs no further significance test.

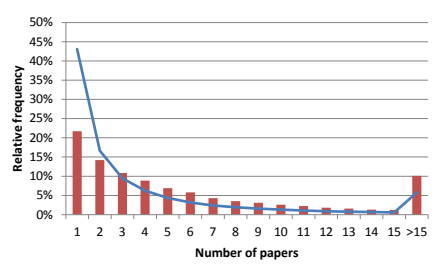


Figure 1. Relative frequency of publication activity of RID authors (bars) vs. all authors (line). [Data sourced from Thomson Reuters Web of Knowledge]

Conclusions

The validity of name disambiguation for some countries like China proved to be beyond tolerance. Nevertheless, the results leave no doubt. The extent of RID registration is still low and differs among countries. We also found that authors with RID are usually more productive than others. RID might therefore not (yet) be used to derive reference standards for publication activity.

Reference

Strotman, A. & Zhao, D. (2012), Author name disambiguation: What difference does it make in author-based citation analysis? *JASIST*, 63(9), 1820–1833.

ARE THE BRIC AND MITS COUNTRIES IMPROVING THEIR PRESENCE IN THE INTERNATIONAL SCIENCE?

Elba Mauleon[1], Daniela De Filippo[2]

[1] *elba114@hotmail.com*
Postdoctoral Fellowship supported by the Spanish Ministry of Education through *FECYT*

[2] *dfilippo@bib.uc3m.es*
LEMI (Department of Library and Information Science), Carlos III University of Madrid, Madrid 126, Getafe, Spain

Introduction

In 2001, the Goldman Sachs group coined the name of BRIC countries to group under this abbreviation the more dynamic emergent economies of the international market: Brazil, Russia, India and China (Wilson and Purushothaman, 2003). Ten years later, the same group proposes a new abbreviation, the MIST to talk about the economies of Mexico, Indonesia, South Korea and Turkey.

Nowadays BRIC and MIST countries are recognized as the most dynamic countries, which are facing well the global crisis. An interesting question is whether this growth can also be evidenced in the scientific and technological fields. We would like to explore the ability of these countries to integrate into the international scientific community and strengthen their positions in the coming years.

Purpose of research

According with this, the objective of this work is to determine if these emergent countries show the same push in the scientific and technological dimensions than the one described in their economic activity. In this sense, the goals of this investigation are:

- To explore the evolution of the scientific output of BRIC and MIST countries in the two last decades. What was the evolution of the world share of BRICS and MITS scientific output?
- To measure the impact of the scientific output of these countries in the international scientific community.

Finally, this work aims at offering data to detect if BRIC and MIST countries are the new strategic actors that will replace the countries that traditionally have led the scientific and technological activity. In this sense, the following indicators were calculated for each country:

a) Scientific activity indicators:

- Annual evolution of the number of documents
- Productivity: scientific production per million inhabitants

b) Influence and impact of the knowledge produced by BRIC and MIST countries:

- Annual evolution of the number of citations per document.
- Percentage of non-cited documents.

c) Socioeconomic indicators:

- Annual evolution of the Gross Domestic Expenditure on Research and Development.

d) International scientific influence of knowledge producing and diffusing institutions:

- Annual evolution of the number of journals from these countries included in JCR.
- Collaboration habits: percentage of documents in national, international and without collaboration.
- Number of academic institutions from BRIC and MITS countries in the international ranks.

e) Specialization Index: the percentage of documents and their annual evolution by broad scientific area was calculated.

Methods

One of the most visible results of the scientific activity is the diffusion of the investigation in peer review journals of recognized prestige. In agreement with this premise, the scientific publications of BRIC and MIST countries have been obtained using as source of data the international database Web of Science (WoS). The use of this database allows us to obtain the scientific production of these countries in its more international scope, which turns out essential to analyze the integration of these countries into the international scientific community.

Another aspect considered in this work has been the presence of higher education institutions from these countries in international rankings, such as the ARWU ranking -elaborated by the University of Shanghai- and THE -produced by TIMEs-.

Finally additional sources –as UNESCO, World Bank- were consulted to obtain data related to other R&D activities. The period of analysis was 1990-2010.

Table 1. Number of journals in JCR by countries

COUNTRY	SCI			SSCI		
	1997	2004	2010	1997	2004	2010
Brazil	9	16	89	2	2	20
Russia	96	104	147	9	8	6
India	37	47	94	5	3	5
China	20	71	138	3	3	6
BRIC	**162**	**238**	**468**	**19**	**16**	**37**
%BRIC vs JCR	3,26	3,99	5,8	1,14	0,93	1,35
Mexico	5	7	28	4	4	13
Indonesia	0	0	0	0	0	0
Turkey	1	3	49	1	1	12
South Korea	6	27	75	0	2	12
MITS	**12**	**37**	**152**	**5**	**7**	**25**
%MITS Total JCR	0,24	0,62	1,88	0,30	0,41	0,91
Germany	381	427	545	52	52	110
Canada	73	75	94	31	28	26
USA	1915	2289	2724	1010	982	1229
France	153	143	189	23	17	25
Italy	51	65	121	1	1	13
Japan	133	160	207	9	7	8
UK	999	1267	1592	325	419	725
G7	3705	4426	5472	1451	1506	2136
%G7 vs JCR	74,65	74,15	67,78	86,78	87,97	78,21
Journals in JCR	4963	5969	8073	1672	1712	2731

Preliminary results

Scientific production

From 1990 to the present time a remarkable growth in the number of publications of the BRIC countries is observed. Its scientific production in WoS increased from 69,764 documents in 1990 (7% of the world) up to 311,077 (17% of the world) in 2010. Of the four, China and Brazil present the greater growth in the period. On the other hand, the scientific production of the MIST represent 5,020 publications in 1990 (0,5% of WoS) and 93,169 in 2010 (5% of the world). These data show the increasing international visibility they have gained. The significant growth of production can be explained, in part, by the increasing number of journals

published in these countries and included in JCR. Table 1 shows the evolution of the number of journals and the percentage each group represents in the total world. It can be seen that while the G7 is still the predominant group, emerging countries have been gaining ground.

Academic visibility

The presence of higher education institutions of BRIC and MIST countries in international rankings is also a reflection of their increasing visibility (Table 2). In less than a decade, not only the number of universities among the world's leading has increased, but also the position of those already included has improved.

Table 2. Presence of BRIC and MIST universities in international rankings

	ARWU		THE	
Country	**2003 (500 univ)**	**2010 (500 univ)**	**2004 (200 univ)**	**2010 (200 univ)**
Brazil	4	7	0	2
Russia	0	2	1	2
India	1	2	0	1
China	12	29	6	16
Mexico	1	1	0	0
Indonesia	Without data		0	0
Turkey	1	1	0	0
South Korea	8	10	3	4

Discussion and conclusions

Throughout this study we have observed a remarkable growth of the BRIC and MITS countries not only through economic indicators but also in the scientific and academic fields. The notable increase of international production, the upward trend in the impact and visibility of these publications and the growing presence of their universities in the international rankings suggests these "developing economies" are evolving towards "knowledge economies" (Dahlman and Aubert, 2001). We consider that these countries are the new players that will have a predominant role in science and technology in the coming decades.

References

Cooper, J (2006) "Of BRICs and brains: Comparing Russia with China, India, and other populous emerging economies" EURASIAN GEOGRAPHY AND ECONOMICS, 47 (3): 255-284

Dahlman, C.y Aubert,J.E China and the Knowledge Economy. Seizing the 21st Century. Washington, DC: World Bank, WDI Development Studies, 2001.

UNESCO (2010) El Informe de la UNESCO sobre la Ciencia, 2010 UNESCO, Paris www.unesco.org/science/psd

Wagner, C; Kit Wong, S (2012) "Unseen science? Representation of BRICs in global science" Scientometrics 90 (3): 1001-1013

Wilson, D. y Purushothaman, R. Dreaming with BRICs: The Path to 2050. New York, NY: Goldman Sachs, Global Economics Paper No. 99, October 1, 2003.

JOURNAL IMPACT FACTOR, EIGENFACTOR, JOURNAL INFLUENCE AND ARTICLE INFLUENCE

Chia-Lin Chang[1], Michael McAleer[2] and Les Oxley[3]

[1] *changchialin@nchu.edu.tw*
Department of Applied Economics, National Chung-Hsing University (Taiwan)

[2] *m.mcaleer@gmail.com*
Econometrisch Instituut, Faculteit der Economische Wetenschappen, Erasmus Universiteit Rotterdam (The Netherlands)

[3] *loxley@waikato.ac.nz*
Department of Economics, University of Waikato, Private Bag 3105, Hamilton (New Zealand)

1. Introduction

We examine the practical usefulness of two new journal performance metrics, namely the *Eigenfactor Score*, which measures "importance", and *Article Influence Score*, which measures "prestige", using ISI data for 2009 for the 200 most highly cited journals in each of the Sciences and Social Sciences. We compare them with two existing ISI metrics, *Total Citations* and the 5-year Impact Factor (5YIF) of a journal. We show that the Sciences and Social Sciences are different in terms of the strength of the relationship of journal performance metrics, although the actual relationships are very similar. Moreover, the importance and prestige journal performance metrics are shown to be closely related to the two existing ISI metrics, and hence add little in practical usefulness to what is already known.

2. Key Research Assessment Measures (RAM).

Leading journal performance measures are:

(1) 2-year impact factor (2YIF); (2) 5-year impact factor (5YIF); (3) Eigenfactor score: The Eigenfactor score (see Bergstrom (2007), Bergstrom, West and Wiseman (2008), and Bergstrom and West (2008)) is a modified 5YIF. **(4) Article Influence:** The Article Influence score measures the relative importance on a per-article basis, and is a standardized Eigenfactor score.

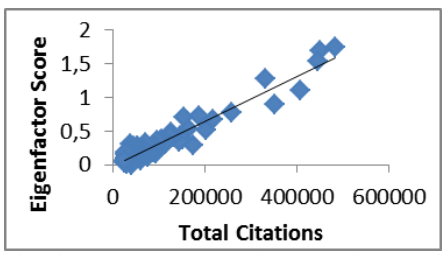
3. Empirical Results

Figures 1-4 evaluate the 200 most highly cited journals, according to 2YIF, in both the sciences and social sciences for 2009. These figures relate the Eigenfactor score to Total Citations and the Article Influence score to 5YIF. The Total Citations data for 2009 for the Sciences and Social Sciences were downloaded from ISI on 19 June 2010 and 20 June 2010, respectively.

A simple linear regression, with the Eigenfactor score as a function of Total Citations, is given in Figures 1 and 3 for the Sciences and Social Sciences, respectively. The estimated model shows that the Eigenfactor score

increases, on average, by 0.000004 and 0.000003 for each unit increase in Total Citations for 2009 for the Sciences and Social Sciences, respectively.

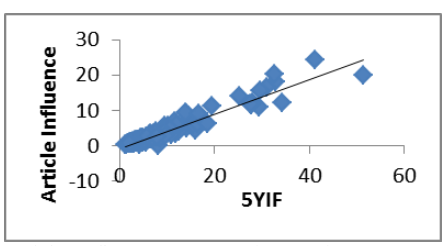
Figure 1. Eigenfactor Score and Total Citations for 200 Most Highly Cited Journals in Sciences for 2009



Eigenfactor Score = -0.022*+3.3E-06* x Total Citations + error, R^2=0.931, * significant at 5%

The goodness-of-fit measures, namely R^2 = 0.931 and R^2 = 0.659 for the Sciences and Social Sciences, respectively, show that the Eigenfactor score can be estimated accurately, especially for the Sciences, on the basis of a simple linear regression against Total Citations.

Figure 2. Article Influence Score and 5YIF for 200 Most Highly Cited Journals in Sciences for 2009



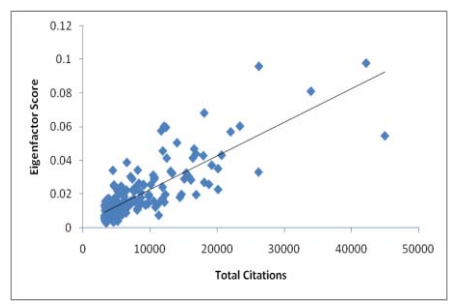
Article Influence = -0.719*+0.489* x 5YIF + error, R^2=0.923, * significant at 5%

The approximate relationships between the Eigenfactor score and Total Citations can be expressed as:

Eigenfactor score = k (Total Citations)

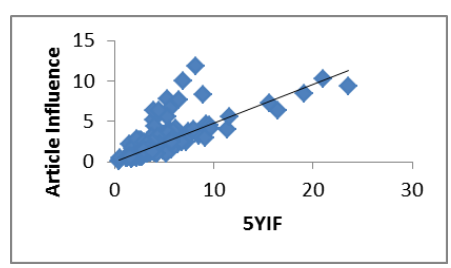
where k = 0.0000033 and k = 0.000002 for Sciences and Social Sciences, respectively. The estimated value of k = 0.00000396 in Ferscht (2009) for the Sciences, based on ISI Total Citations data for 2007, is in accordance with the result obtained in the present paper, as is the value of R^2**.** Another simple linear regression, with the Article Influence score as a function of 5YIF, is given in Figures 2 and 4 for 2009 for the Sciences and Social Sciences, respectively. The estimated models show that the Article Influence score increases, on average, by 0.489 and 0.479 for each unit increase in 5YIF for 2009 for the Sciences and Social Sciences, respectively.

Figure 3. Eigenfactor Score and Total Citations for 200 Most Highly Cited Journals in Sciences for 2009



Eigenfactor Score=0.029*+1.99E-06* x Total Citations + error, R^2=0.659, * significant at 5%

Figure 4. Article Influence Score and 5YIF for 200 Most Highly Cited Journals in Sciences for 2009



Article Influence=0.160*+0.479* x 5YIF + error, R^2=0.572, * significant at 5%

The goodness-of-fit measures, as given by R^2 = 0.923 and R^2 = 0.572 for 2009

for the Sciences and Social Sciences, respectively, show that the Article Influence score can be approximated very accurately for the Sciences, and reasonably accurately for the Social Sciences, on the basis of a simple linear regression relationship of Article Influence score against 5YIF, namely:

Article Influence score = 5YIF/2.

Although the goodness-of-fit value of R^2 obtained in the present paper is slightly higher than in Franceschet (2009), namely $R^2 = 0.880$, in relating the Article Influence score to 5YIF, the latter paper had an effect of 5YIF on Article Influence score of 0.452, which is very similar to that proposed above.

4. Conclusions

Although the Sciences and Social Sciences are dramatically different in terms of the strength of the underlying relationship of the journal performance metrics considered in this paper, the actual empirical relationships are very similar. As Article Influence is a modification of 5YIF, it is perhaps not surprising that the two scores are highly and positively correlated.

Given the very high correlations between the Eigenfactor score and Total Citations, and between the Article Influence score and 5YIF, and the corresponding high R^2 values for the simple linear regressions, the Eigenfactor score and Article Influence score would not seem to be entirely necessary for the Social Sciences, and not at all necessary for the Sciences, relative to the leading journal performance measures that are already available, namely Total Citations and 5YIF, respectively.

5. References

Arendt, J. (2010), Are article influence scores comparable across scientific fields?, *Issues in Science and Technology Librarianship*, 60.

Bergstrom C. (2007), Eigenfactor: Measuring the value and prestige of scholarly journals, *C&RL News*, 68, 314-316.

Bergstrom, C.T. and. West (2008), Assessing citations with the Eigenfactor™ metrics, *Neurology*, 71, 1850–1851.

Bergstrom, C.T., J.D. West and M.A. Wiseman (2008), The Eigenfactor™ metrics, *Journal of Neuroscience*, 28(45), 11433–11434.

Davis, P.M. (2008), Eigenfactor: Does the principle of repeated improvement result in better estimates than raw citation counts?, *Journal of the American Society for Information Science and Technology*, 59(13), 2186-2188.

Elkins, M.R., C.G. Maher, R.D. Herbert, A.M. Moseley and C. Sherrington (2010), Correlation between the journal impact factor and three other journal citation indices, *Scientometrics*, 85, 81-93.

Fersht, A. (2009), The most influential journals: Impact factor and Eigenfactor, *Proceedings of the National Academy of Sciences of the United States of America,* 6883-6884.

Franceschet, M. (2010), Journal influence factors, *Journal of Informetrics*, 4, 239-248.

ISI Web of Science (2010), Journal Citation Reports, Essential Science Indicators, Thomson Reuters ISI.

Rousseau, R. et al. (2009), On the relation between the WoS impact factor, the Eigenfactor, the SCImago journal rank, the article influence score and the journal h-index, *Conference Proceedings*, Nanjing University, 2009.

ASEP ANALYTICS. A SOURCE FOR EVALUATION AT THE ACADEMY OF SCIENCES OF THE CR

Jana Doleželová[1] and Zdeňka Chmelařová[2]

[1]*dolezelova@knav.cz*, [2]*chmelarova@knav.cz*
Library of the Academy of Sciences of the Czech Republic v.v.i., Národní 3, 115 22 Prague, Czech Republic

Repository of the Academy of Sciences of the Czech Republic [1]

Since 1994, the **Library of the Academy of Sciences of the Czech Republic** [2] has been the coordinator of bibliographic database ASEP, which contains the records of publishing activities of 54 institutes of the Academy of Sciences of the Czech Republic. The total number of records exceeds 216,000 (with an average annual increase of 11,000), they are divided into 29 categories, and 72 trained administrators of institutional data participate in their creation. Starting from 2012, full texts may be saved with each record. The data is published as an on-line catalogue, selected entries can be displayed in different formats, and the data may be printed out or saved for future use. This database is used to evaluate the scientific results achieved by the institutes, departments or individual researchers of the Academy of Sciences of the Czech Republic, either in that institution or in the entire country. Within the Czech Republic, the evaluation is based on the results submitted to Information Register of Research and Development results [3], maintained by the Research, Development and Innovation Council [4].

Fig. 1: Home page of ASEP Analytics for an institute

ASEP Analytics [5]

ASEP Analytics was created as a software extension that provides analytical reports derived by a combination of queries and calculations from the data stored in the ASEP database, which cannot be displayed directly in the catalogue. Each institute, its scientific teams and authors have access to web pages with the same graphic layout - **menu, filters** for setting the query (time period, document type etc.) and the **field of data** (Fig. 1).

Graphic Presentations in ASEP Analytics

Analyses of research areas

Institutes of the **Academy of Sciences of the Czech Republic** are active in three distinct research areas (I. Mathematics, Physics and Earth Sciences, II. Life and Chemical Sciences, III. Humanities and Social

Sciences), which are further divided into sections.
Examples of charts for research areas (Fig. 2) and charts for sections (Fig. 3).

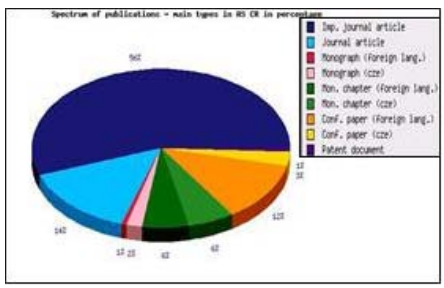


Fig. 2: Number of documents from the Academy of Sciences of the CR for a selected period of time.

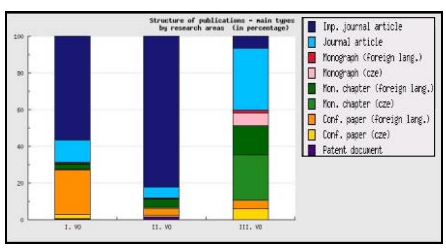


Fig. 3: Number of documents in sections for a selected period of time.

Analyses of institutes, scientific teams (departments), and researchers

a) Summaries of publications by institutes or researchers in a selected period of time. Every record shows a reference to the Information Register of Research and Development results, which is a characterization of the result within the Czech Republic, a link to Web of Science database, DOI, R&D Council evaluation, or possibly a full text in the repository (Fig. 4).

b) Statistics by Institute/Department/ Author
The charts and tables can display a variety of indicators such as the number of results pertaining to specific types of documents, or journals with impact factor, or applied outputs (patents, pilot plant, prototype, specialized map, industrial and utility model, methodology, norms, software) for a selected period of time. It is also possible to obtain diverse outputs and statistics, e.g. the average impact of an institute (Fig. 5) or the number of different types of results for each department within an institute (Fig. 6).

0360530 - MU-W 2012 RIV CZ eng J - Článek v odborném periodiku
Brandts, J. - Korotov, S. - Křížek, Michal[6]
Generalization of the Zlámal condition for simplicial finite elements in Rd.
Applications of Mathematics. Roč. 56, č. 4 (2011), s. 417-424. ISSN 0862-7940
Grant CEP: GA AV ČR(CZ) IAA100190803
Institutional Research Plans: AV0Z10190503
Keywords: linear finite element * mesh regularity * minimum angle condition * convergence
Kód oboru RIV: BA - Obecná matematika
Impakt faktor: 0.480, rok: 2011
http://www.springerlink.com/content/w6433tku3l3x032u/
DOI: 10.1007/s10492-011-0024-1
WOS
Výsledek v RIV
Hodnocení v RIV 2012
ASEP - Institucionální repozitář AV ČR

Fig. 4: Record result (journal with impact factor article)

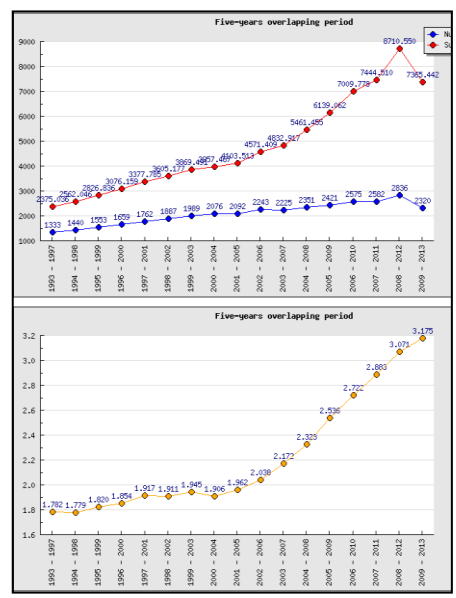


Fig. 5: Average impact factor of an institute

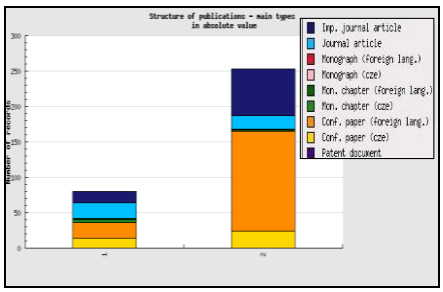


Fig. 6: Number of results by document type and department

c) Relevance of financial support

The results are obtained with financial support from the government budget and other financial sources. It is possible to show the results of the institutes for projects funded from the public budget, which appear in the Central register of Research and Development projects database [6] administered by the Research, Development and Innovation Council, inclusive of their rating by Council methodology, along with the results arising from institutional funding.

Conclusion

ASEP Analytics displays summaries and statistical data showing the results attained by the individual institutes of Academy of Sciences of the Czech Republic. Evaluation of scientific results is a complex task, and the information, especially for a qualitative evaluation by ASEP Analytics, is being gradually expanded and perfected as required by the rules of evaluation in the Czech Republic. ASEP Analytics generates synoptic information that constitutes one of the tools for evaluation, particularly qualitative, of institutes and researchers by various subjects.

Acknowledgement

Project was supported with institutional support RVO:67985971.

References

[1] http://www.library.sk/i2/i2.entry.cls?ictx=cav&language=2&op=search
[2] http://www.lib.cas.cz/
[3] http://www.isvav.cz/prepareResultForm.do
[4] http://www.vyzkum.cz
[5] http://www.lib.cas.cz/arl/
[6] http://www.isvav.cz/prepareProjectForm.do

ASSESSING AN INTERVAL OF CONFIDENCE TO COMPILE TIME-DEPENDENT PATENT INDICATORS IN NANOTECHNOLOGY

Douglas H. Milanez[1], Thiago D. Macedo[2], Roniberto M. Amaral[3], Leandro I. L. de Faria[4] and Jose A. R. Gregolin[5]

[1] *douglas@nit.ufscar.br*
Federal University of Sao Carlos, Washington Luis Highway, km 235, São Carlos – SP (Brazil)

[2] *thiagoma89@gmail.com*
Federal University of Sao Carlos, Washington Luis Highway, km 235, São Carlos – SP (Brazil)

[3] *roniberto@nit.ufscar.br*
Federal University of Sao Carlos, Washington Luis Highway, km 235, São Carlos – SP (Brazil)

[4] *leandro@nit.ufscar.br*
Federal University of Sao Carlos, Washington Luis Highway, km 235, São Carlos – SP (Brazil)

[5] *gregolin@nit.ufscar.br*
Federal University of Sao Carlos, Washington Luis Highway, km 235, São Carlos – SP (Brazil)

Introduction

Patent indicators have been extensively used to assess nanotechnology developments due to their objectiveness and great capability to compare results in complex and interdisciplinary fields. Patent documents, which include patent application and granted patent, are rich sources information and they can be used to depict trends and support planning and decision making (Mogee, 1997; OECD, 2009). Usually, the bibliographic data from patent documents are used to develop patent indicators, thus bibliographic databases, such as the worldwide Derwent Innovations Index (DII) has been preferably used (Wang & Guan, 2012; Milanez, Morato, Faria & Gregolin, 2013). An important criterion for compiling patent indicators is the reference year because every patent document includes several dates reflecting the timing of the invention, the patenting process and the strategy of the applicant. The OECD Patent Manual (2009) recommends using the priority date as it is closest to the date of invention. Nevertheless, there is a delay between the date of application and the date when the information becomes available in databases. This delay includes the period until the publication of the patent application, which varies according to national regulations, but it is usually assumed to be 18 months after the filing date (Mogee, 1997; OECD,

2009) and the period for index bibliographic data. Consequently, it can be observed that there have been declines in late years of time-dependent indicators. This paper aims to analyse the number of patents available in DII according to the interval between the year of the first applications and the year of indexing in order to establish an interval of confidence to compile patent indicators in nanotechnology.

Materials and Method

Samples of nanotechnology and some selected nanomaterial patent bibliographic data were retrieved from the DII using the search expressions from Table 1 applied in the Topic field. In the case of nanotechnology, it was used the modular search strategy proposed by Porter et al (2008). All searches were carried out between 20 and 21 May, 2012 for the time spam from 2000 to 2011 and the total of patent documents recovered is shown in Table 1. All data were collected and analyzed separately using the bibliometric software VantagePoint® (version 5.0).

Table 1. Total of patent documents for each subject analyzed.

Subject	Total of Patent
Nanotechnology	161,121
"carbon nanotub"*	16,094
*fulleren**	3,071
*graphene**	1,897

The delay period for each record was investigated comparing the interval from the year of the prioritizing (first filing) and the year of indexing the patent bibliographic data in DII, which was accessed from the Derwent Primary Accession Numbers. According to DII (2013), this code consists of the year when items entered in the database followed by a six digit code. For every delay period, the number of patents and their percentage representation was obtained. For instance, 23,758 nanotechnology patents were indexed in 2010; from this total, 8,928 were applied firstly in 2009, which means a two year delay and represents 37.6% from the total nanotechnology patent indexed in 2010. After calculating the percentage for every year of indexing, the average was obtained for each delay period. Moreover, the general average was calculated for all subjects.

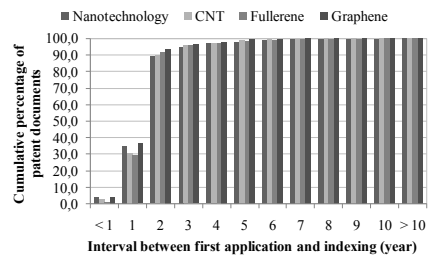


Figure 1. Cumulative percentage of patent documents available in database according to the delay period.

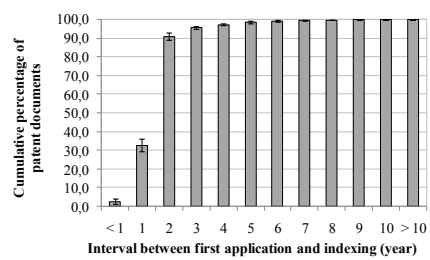


Figure 2. Average of the cumulative percentage of patent documents available in databases according to the delay period.

Results and Discussion

The cumulative percentage of patent documents available in databases increased stably for general (nanotechnology) and specific (nanomaterials) subjects, as can be seen in Figure 1. This result suggests the

database indexes in similar criteria regardless of the subject or its size in terms of the number of patents. Figure 2 presents the average cumulative percentage of the patent documents from all subjects evaluated according to the delay period.

Only 2.68% of the whole patent dataset was available on the DII database less than a year from the first application whereas this value increased to just 32.9% after one year. However, the cumulative percentage grew sharply to 90.9% after two years from the application and it is a consequence for the general patent secrecy of 18 months. Furthermore, as suggested by the charts, some patent documents could take more than 10 years before being indexed on the DII database. A possible explanation for this might be that the database does not monitor all patent repositories from worldwide; therefore a document from a non-monitored country may enter in the database only if it is published in a monitored country. The availability of patent documents in DII to compile patent indicators increased when the interval of the first application and the indexing period is high enough. This means that for a search conducted for the time spam (indexing year) from 2000 to 2011, few patents from the priority year of 2011 and 2010 will be accessible. Even for patent documents from 2009, there is a lack of some documents, although it has given confidence to state indicators until this year.

Conclusion

The delay between the first application and the indexing period in DII is mainly due to the regular period of 18 months of secrecy in most countries' regulation. Therefore, most patent documents are accessible in the database after at least two years. The same time the outcomes set the availability of the patent documents, they also give intervals of confidence to develop time-dependent indicators and other quantitative analyses.

Acknowledgements

The authors are grateful to the Brazilian National Council for Technological and Scientific Development (process number 160087/2011-2), the São Paulo Research Foundation (process number 2012/16573-7) and the Graduate Program in Materials Science and Engineering at the Federal University of São Carlos.

References

Milanez, D. H., Morato, R., Faria, L. I. L., & Gregolin, J. A. R. (2013). Assessing nanocellulose developments using science and technology indicators. *Materials Research*. Retrieved March 15, 2013 from http://www.scielo.br/pdf/mr/2013nahead/aop_1715-12.pdf

Mogee, M. E. (1997). Patents and Technology Intelligence. *Keeping abreast of science and technology: technical intelligence for business*. Columbus: Battelle Press.

OECD. (2009). *OECD Patent Statistics Manual*. Retrieved March 10, 2013 from http://dx.doi.org/10.1787/9789264056442-en

Porter, A. L., Youtie, J., Shapira, P., & Schoeneck, D. J. (2008). Refining search terms for nanotechnology. *Journal of Nanoparticle Research*, *10*(5), 715–728.

Wang, G., & Guan, J. (2012). Value chain of nanotechnology: a comparative study of some major players. *Journal of Nanoparticle Research*, *14*(2).

Derwent Innovations Index. (2013). *Glossary of Thomson Scientific terminology*. Retrieved March 15,

2013 from http://ip-science.thomsonreuters.com/support/patents/patinf/terms/#P

BIBLIOMETRIC INDICATORS OF YOUNG AUTHORS IN ASTROPHYSICS: CAN LATER STARS BE PREDICTED?

Frank Havemann[1] and Birger Larsen[2]

[2] *frank.havemann@ibi.hu-berlin.de*
Institut für Bibliotheks- und Informationswissenschaft, Humboldt-Universität zu Berlin, Dorotheenstraße 26, D 10099 Berlin (Germany)

[2] *blar@iva.dk*
Royal School of Library and Information Science, Birketinget 6, DK-2300 Copenhagen S (Denmark)

Introduction

Most bibliometric indicators are not developed for the evaluation of individual researchers. We test selected indicators with respect to their validity at the level of the individual researcher by estimating their power to predict later successful researchers. For this reason, we compare bibliometric indicators of a sample of astrophysics researchers who later co-authored highly cited papers (later stars, for short) before their first landmark paper with the distributions of these indicators over a random control group of young authors in astronomy and astrophysics.

Here we present results obtained with some standard basic indicators (Wildgaard et al. 2013). We will extend the study to more sophisticated measures with the aim to find the best indicators for predicting later stars. We imagine that later stars apply for a job in an astrophysical research institute five years after their first paper in a journal indexed in Web of Science (WoS). Do they perform better bibliometrically than the average of applicants with the same period of publishing?

Data and method

We inspected 64 astronomy and astrophysics journals to find researchers who started publishing after 1990 and had published for a period of at least five years in WoS journals. We excluded those who had more than 50 co-authors on average because evaluating those big-science authors cannot be supported by bibliometrics. We draw a random sample of 331 authors mainly publishing in this field and affiliated longer in Europe then elsewhere. The latter criterion contradicts with the international character of astrophysics research but makes the sample more homogenous with respect to the educational and cultural background of the researchers.

To find authors with highly cited papers, for each journal considered we ranked papers with more than four citations per year and less than ten authors according to their citations per year. We excluded papers with ten or more authors because we want to have later stars whose contributions to the successful papers are not to small. From the top 20 percent of these paper rank-lists we extracted all European authors of highly cited papers. We obtained 362 candidates who

published their first highly cited paper at least five years after their first paper in one the 64 journals.

We ranked these later-star candidates according to their number of highly cited papers. We went through this list and checked whether the authors had really five years or more to wait for the break-through paper if all their papers in WoS-journals are taken into account. We chose the first 40 authors to keep the effort manageable.

For all WoS-papers of the 40 later stars and of the 331 random authors (downloaded at Humboldt-University, Berlin) all citing papers were determined by CWTS, Leiden.

All bibliometric indicators presented below are based on papers and their citations within the first five years of the author. To compare only authors with similar collaboration behaviour we restricted both samples to authors with less than four and more than one co-author on average ending up with 30 later stars and 179 random authors.

For each bibliometric indicator considered, we test whether both samples behave like random samples drawn from the same population by applying a one-sided Wilcoxon rank sum test with continuity correction. We test the null hypothesis that for both samples we have the same probability of drawing an author with a larger value in the other sample. The alternative hypothesis is that indicator values of later stars exceed the values of random authors (cf. wikipedia http://en.wikipedia.org/wiki/Mann-Whitney-Wilcoxon_test).

Results

Until now we have calculated and tested two absolute output and two absolute influence indicators, respectively:

number of papers,

fractional paper score giving each of k authors of a paper a $1/k$-fraction of it,

number of citations, and

fractional citation score giving each of k authors of a paper a $1/k$-fraction of its citations.

In addition we calculated the widely used Hirsch index, a number combining influence and output performance in an uncontrolled manner (Hirsch 2005). The later stars perform somewhat better than random authors (s. Table 1) but the distributions are rather similar (s. Figures 1 and 2).

In the last column of Table 1 we list the failure probability p of rejecting the null hypothesis that both samples behave like random samples drawn from the same population.

Table 1: Median indicators of samples and test probability *p*

Indicator	*stars*	*random*	*p*
Number of papers	8	6	.083
Fractional score	2.72	2.00	.062
Nr. of citations	38	25	.051
Fractional citations	10.51	6.67	.031
Hirsch index	3	3	.245

Discussion

The high failure probabilities of rejecting the null hypothesis indicate that the differences found are not statistically significant for all indicators but the fractional citation score where we have at least a significance on the 5 percent level. Thus, it is very unlikely to discover a later star in astrophysics by comparing her output with the output of a random author. The Hirsch index makes no difference at all. The number of citations does also not help much. There is also no difference when we compare citations per paper and year of both samples (data not shown). The only moderately helpful of the indicators considered here is fractionally counted

number of citations. We will extend the study to other indicators of output and influence including variants of the Hirsch index.

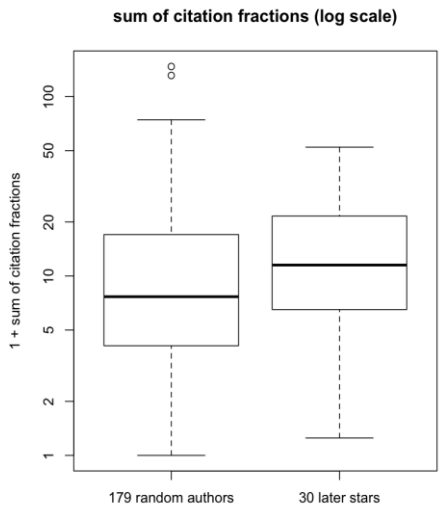


Figure 2. Sum of citation fractions

Acknowledgments

We thank Jesper Schneider for helpful discussions and Paul Wouters for providing citation data. The analysis was done for the purposes of the ACUMEN project, financed by the European Commission, cf. http://research-acumen.eu/.

References

Hirsch, J. E. (2005). An index to quantify an individual's scientific research output. *Proceedings of the National Academy of Sciences* 102(46), p. 16569–16572.

Wildgaard, L., Schneider, J., & Larsen, B. (2013). Quantitative Evaluation of the Individual Researcher: a review of the characteristics of 114 bibliometric indicators. *Forthcoming*.

BIOLOGICAL SCIENCES PRODUCTION: A COMPARATIVE STUDY ON THE MODALITIES OF FULL PHD IN BRAZIL OR ABROAD

Daniel Henrique Roos, Luciana Calabró, Nilda Vargas Barbosa, João Batista Teixeirada Rocha and Diogo Onofre Souza

daniel_varzeano@hotmail.com; luciana.calabro.berti@gmail.com; nvbarbosa@yahoo.com.br; jbtrocha@yahoo.com.br; diogo@ufrgs.br
Federal University of Rio Grande do Sul, Departamento de Bioquímica, Rua Ramiro Barcelos, 2600 – anexo, Porto Alegre, RS (Brazil)

Introduction

Science has been regarded as an information production system especially in the form of publications. Therefore, science must be considered as a broad social system whose function is crucial to promote the dissemination of scientific and technological knowledge [1].

The academic structure built in Brazil allowed a significant expansion of the national scientific community and its scientific production. Until the mid-1990s, the construction of this scientific base occurred relatively fast by the training of masters and doctors abroad, with a legal commitment to return to Brazil. Three mechanisms enabled this policy: 1- the awarding of grants to professionals with a permanent position in an institution in Brazil, 2- the inclusion of clauses specifying the immediate return after getting the title in the deed of undertaking signed by fellows; 3- the government efforts for the establishment of international agreements with "receiving countries" to prevent the granting of a residence permited to former fellows [2]. Therefore, this work was planned to analyze and compare the evolution of scientific production (number of articles published, number of citations, authoring type and journal impact factor (IF)) of researchers with full doctorate in Brazil or abroad in the areas of biochemistry, biophysics, physiology, pharmacology and related fields.

Methodology

The present study compared the scientific production of former fellows with Full Doctorade in Brazil (FDB) or Abroad (FDA) in the area classified by CAPES as Biological Sciences II (Physiology, Biochemistry, Pharmacology and related areas such as molecular biology, cell biology, Enzymology, etc). The scientific production of researchers was analyzed during the first nine (9) years after the end of doctorate course and covered a sample (total along the period) of 19 (nineteen) former fellows from each modality. The data collected included: number of articles published; number of citations and impact factor of the journal. These data were obtained from the database of CAPES, CNPq Lattes, Web of Science (Institute for Scientific Information (ISI) and Scival - Scopus).

Results

The number of articles published by FDB increased gradually in the three

triennia analyzed after the end of doctorate (Fig. 1). This growth profile was not observed for FDA. The results also demonstrate that the average of articles published by FDB was higher than FDA during the last two triennia (Fig. 1).

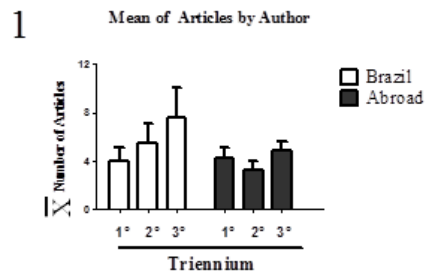


Fig. 1- Mean of articles published by triennium. Open bar; Brazilian former fellows and Closed bar; Abroad former fellows. n = 19.

The IF and number of citations were analyzed in order to evaluate qualitative parameters of publications. The mean IF of articles published by FDB is approximately 2.0 (two) and that this value remain constant during the studied period (Fig. 2). The mean IF of article from FDA was also constant during the analyzed triennia; however with a higher value (around 3.5 – 4) than that from FDB (Fig. 2).

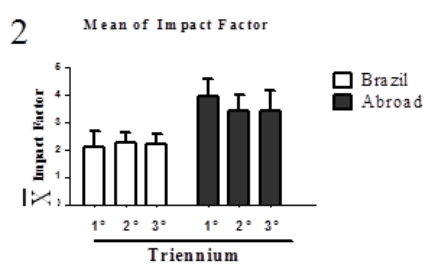


Fig. 2 - Mean of impact factor by triennium. Open bar; Brazilian former fellows and Closed bar; Abroad former fellows. n = 19.

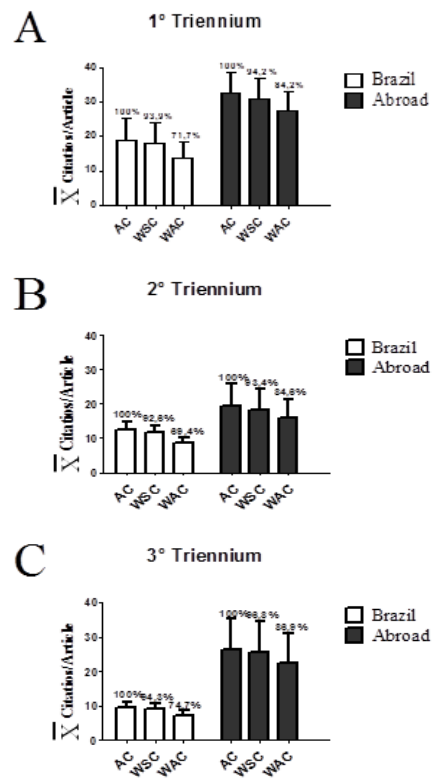


Fig. 3- Mean of citations per article by triennium. First triennium 3A; Second triennium 3B; Third triennium 3C; AC – All Citation; WSC – Without Self Citation; WAC Without All Citation of Authors; Open bar; Brazilian former fellows and Closed bar; Abroad former fellows. n = 19.

The average of citations by articles in the respective triennia for FDA and FDB is shown in Figure 3. In order to detail the analysis regarding the quality of articles published, the data were also evaluated without self-citations (WSC: without the self-citations of the author and WAC: without the citations of all authors of the article). The results show that in all triennia, the average of citations by article is higher for FDA when compared to the FDB. Indeed, the percentage of self-citation WAC in the

articles published by FDB is slightly larger than FDA (Fig. 3). Taken together, the data depicted in Figure 2 and 3 indicate a higher impact of articles published by FDA than that of FDB.

The number of articles published as corresponding author may be considered, at least in part, an indicator of scientific independence of the ex-fellows, since the corresponding author is expected to answer questions of reviewers and interacts with the journal and the scientific community. Moreover, in most of the cases, the corresponding author is the person responsible for the scientific project and usually the scientific advisor of one of the other authors. To include in the work a measure of scientific independence, we analyzed the percentage of articles published by FDB and FDA as corresponding author. For both doctorate modalities the percentage of articles as corresponding author increases during the triennia period. This parameter was slightly larger for FDB when compared to FDA (Fig. 4). Overall, these results indicate that the scientific independence of FDB is greater and faster than FDA.

Conclusions

The results showed in this study allow us to infer that, in terms of number, the FDB may be contributing significantly to the current growth in the quantity of scientific articles published by Brazilian community. However, the quality of articles published by FDB was lower than that published by FDA (quality here evaluated both by the IF of the journals and by the number of citation of the papers), perhaps, indicating that the production of scientific articles by FDB was more centered in quantity than in quality. On the contrary, FDA publication attitude seems to be more directed to quality than quantity. In a general, the results presented here signalize to the government policies the need to invest more in the post-graduation programs to enhance the relevance of Brazilian articles in terms of scientific quality

Acknowledgements

This work was supported by grants from, UFSM; UFRGS; FAPERGS; CAPES; CNPq; FINEP (IBN-Net) and INCT-EM.

References

[1] Macias-Chapula, C.A. (1998). O papel da informetria e da cienciometria e sua perspectiva nacional e internacional. *Ciência da Informação*, *27*(2), 134-140.

[2] Schwartzman, S. (1978). Struggling to be born: The scientific community in Brazil. *Minerva, 16*(4), 545-580.

A CITATION ANALYSIS ON MONOGRAPHS IN THE FIELD OF SCIENTOMETRICS, INFORMETRICS AND BIBLIOMETRICS IN CHINA (1987-2010)

Yuan Junpeng, Yang Yang, Pan Yuntao, Wu Yishan

junpengyuan@126.com
Institute of Scientific and Technical of Information of China, Beijing 100038, China

Introduction

Scholarly books and monographs play a significant role in research communication, providing important scientific review in some disciplines and the latest research for others. Many studies have analysed citations to books (Small 1979, Glänzel 1999, Tang 2008, Giménez-Toledo2009, Kousha, 2009). There are only few investigations have examined citations from books (Chung 1995, Book Citation Index, Chen 2009). This article carries out citation analysis on monographs in the field of scientometrics, informetrics and bibliometrics in China between 1987 and 2010. Using metrology and statistics methods to analysis references in the monographs, paper found some interesting conclusions and partially expanding the object of citation analysis to monographs.

Data and Methods

The data sources of monographs are from the National Library of China (NLC). We accessed monographs database of NLC, use the keywords scientometrics, informetrics and bibliometrics to search monographs. Overall, the search strategy retrieved about 28 volumes books in the field since 1987. After delete 7 volumes symposiums, get 21 volumes monographs, the monographs are ranked by publication year in table 1.

Table1 The bibliographic of scientometrics, informetrics and bibliometrics monographs in China

serial number	Title	Publi-cation year	Author	Age	Gender
1	An Introduction to Bibliometrics	1987	Luo Shisheng	42	Male
2	Bibliometrics	1988	Qiu Junping	41	Male
3	Bibliometrics Course	1990	Wang Chongde	52	Male
4	An Basic of Bibliometrics	1993	Ding Xuedong		Male
5	Generality on Bibliometrics	1994	Luo Shisheng	49	Male
6	Scientometrics: Indicator, Model, Application	1995	Liang Liming	46	Female
7	Methodology of Scientometrics	1999	Pang Jingan	49	Male
8	Indicator of Science and Technology and Evaluation Method	2000	Luo Shisheng	55	Male
9	Development Science and Technology Mathematical Principle	2005	Gu Xingrong	54	Male
10	Scientometrics: Theoretical Exploration and Case Study	2006	Liang Liming & Wu Yishan	57 /48	Female/ Male
11	Methodology Research of Scientometrics	2006	Fang Yong	38	Male
12	Network Information Resources Evaluation	2007	Pang Jingan	55	Male
13	Informetrics	2007	Qiu Junping	60	Male
14	An Introduction to Informetrics	2007	Guo Qiang & Liu Junyou	31	Male
15	Mapping Knowledge	2008	Hou Haiyan	37	Female

	Domains of Scientometrics				
16	Content Analysis in Bibliometrics	2008	Qiu Junping	61	Male
17	Webometrics: a Theoretical and Empirical Study	2009	Zhang Yang	34	Male
18	Webometrics: Theory, Tools and Applications	2009	Sun Jianjun & Li Jiang	47/27	Male
19	Informetrics and Application in Medical	2010	Wang Wei	51	Male
20	Advanced Course in Scientometrics	2010	Yuan Junpeng	37	Male
21	Scientometric Analysis of Highly Cited Papers(1979-2008)	2010	He Defang & Zheng Yanning	47/45	Male

Note: Serial number in table 1 on behalf of monographs in follows.

Results

Overview on scientometrics, informetrics and bibliometrics monographs in China

As shown in Table 1, the earliest monograph is An Introduction to Bibliometrics, which was published in 1987 and the last one was published in 2010. With the development of the discipline, the topics of the monographs include bibliometrics, scientometrics, informetrics and webometrics. The number of the monographs in the field kept increasing, especially after 2000, the number is 13 accounting for 62% of the total sample.

Authors analysis

Table 1 shows there are 19 authors who write the monographs in the field. Four monographs are co-author books, representing 19.05% of the total number of monographs in this study, the serial number are 10, 14, 18 and 21. Until 2006, co-author books are appearing and it might be because communication in research is more convenient. Four authors write multiple monographs: Luo Shisheng and Qiu Junping write 3 monographs, followed by Liang Liming and Pang Jingan write 2 monographs. Two authors are female and the other seventeen authors are male. The average age when the author published the monographs is 46. Further analysis of authors' institution indicates that the Chinese universities (13 out of 19) are the majority, while there are only 6 research institutes.

Citations per monograph analysis

The total citation rates of 21 monographs is 3628 and citations per monograph is 173, the top one is 807. It is showed there was no trend for the citations per monograph before 2005, because the distribution of monographs is more dispersed. The number of citations per monograph has been increasing since 2005 and especially after 2008. Further investigation shows that monographs exceeded average references mostly concentrated in scientometrics and informetrics.

References type analysis

The references type of the 21 monographs mainly is journal and monograph, so we divided the references type into three types: journal, monograph and others. Scientific journals as records, reports, dissemination and accumulation of scientific information carrier, while delivering high-quality scientific knowledge, but also to achieve timeliness, breadth, continuity, this cannot be done by other type of scientific literature. The type of journal has 2568 references, representing 70.78% of the total number of references, followed by monograph (582) and others (478). From each monograph, the type journal is more than type monograph in 17 monographs. It is worth mentioning that the type journal is less than type monograph in No.7, 9, 14 monographs, especially No. 14 monograph has only monograph

references. It might be related to the mode of author absorb knowledge and cite the references.

References language branch analysis

Chinese, English, Japanese, Russian and Italian, etc. are the references language. Because the numbers of reference write in Japanese, Russian and Italian is very small we divided the language branch into three types: Chinese, foreign and translation. In terms of the total number of references, foreign language demonstrated its dominant position; Chinese and translation ranked second and third. It indicates that researchers in this field often consult the foreign language literature, tracking the latest research developments, inevitable cite in their own academic achievement as a reference. It is worth mentioning that there only Chinese references in No.8 and 9, and the number of foreign references in No.4 is 133, the number of Chinese references in this monograph is only 10.

Time distribution of references analysis

Analysis the references publication year distribution, we can understand the best time to use literature; evaluate the strength of authors to absorption new information. References have demonstrated a stable and strong growing trend. For the increase rate of references the growth pattern can be divided into four periods. The first period is from 1909 to 1969, in this period, the number of references increased slowly, from 1 in 1909 to 26 in 1969. The second period is from 1970 to 1980, the number of references increased more quickly, from 15 in 1970 to 40 in 1980. The third period is from 1981 to 1995, the number doubled to the second period, but the number is volatility. The fourth period is from 1996 to 2009, both the number and the speed increased dramatically, the number is 215 in 2005. Further analysis shows that the time of citations concentrated is parallel to the time of monographs dense, which is consistent with the literature use laws.

Conclusions

In the field of scientometrics, informetrics and bibliometrics, four monographs are co-author books and four authors write multiple monographs and the average age when the author published the monographs is 46. In addition, the Chinese universities (13 out of 19) are the majority. Furthermore, the number of citations per monograph has been increasing since 2005 and especially after 2008. The study aims to divulge the patterns of monographs' reference, cited references in books can be difficult to locate, so we select a familiar field scientometrics to study. The next stage of this study will invite experts in the field to explain the results from the scientometric analysis.

Acknowledgments

This project was supported by a grant (No.70673019) from the National Science Foundation of China (NSFC).

References

Chen Guangyu(2009). Sample Selection and Citation Analysis: The Case of Monograph in the Research on the Information Sources of US China Study. Library and Information Service 53(14): 20-22, 27. (In Chinese)

Chung, Y. (1995). Characteristics of references in international classification systems literature. Library Quarterly, 65(2), 200-215.

Giménez-Toledo, E. and A. Román-Román (2009). Assessment of humanities and social sciences monographs through their

publishers: a review and a study towards a model of evaluation. Research Evaluation,18(3): 201-213.
Glänzel, W., & Schoepflin, U. (1999). A bibliometric study of reference literature in the sciences and social sciences. Information Processing & Management, 35(1), 31-44.
Kousha, K., & Thelwall, M. (2009). Google book search: Citation analysis for social science and the humanities. Journal of the American Society for Information Science and Technology, 60(8), 1537-1549.
Schubert, A. (2002). The Web of Scientometrics. Scientometrics, 53(1), 3-20.
Small, H. & Crane, D. (1979). Specialties and disciplines in science and social science: An examination of their structure using citation indexes. Scientometrics, 1(5-6), 445-461.
Tang, R. (2008). Citation characteristics and intellectual acceptance of scholarly monographs. College & Research Libraries, 69(4), 356-369.

CITATION PATTERNS FOR SOCIAL SCIENCES AND HUMANITIES PUBLICATIONS

Lucy Amez

lucy.amez@vub.ac.be
Vrije Universiteit Brussel, R&D department and ECOOM
Pleinlaan 2, 1050 Brussel

Introduction

The distribution of research funding provided by the special Research Fund of the Flemish Government (BOF) to universities is based on a set of performance driven indicators (Debackere and Glänzel 2004) where the bibliometric component plays an important role. The majority of the data are collected and monitored by ECOOM, the Centre for R&D Monitoring. Whereas initially only Web of Science (WOS) indexed items were counted, since 2008 bibliometric data are enriched with publications indexed by the Flemish Academic Bibliographic Database for Social Sciences and Humanities (VABB-SHW).

The VABB-SHW is a retrospective bibliographic database, assembling publications by authors attached to social sciences and humanities disciplines at a Flemish University. Publication types include, apart from journal/proceedings articles, author/editor of book and book chapters. The BOF regulation stipulates a number of eligibility criteria for the inclusion of an item, conditions which are both formal (publicly accessible, identification by ISBN/ISSN), as well as content related: the publication must have been submitted to a prior peer review process by experts in the discipline and it must contribute to the development of new insights or applications. An Authoritative Panel (AP) of 18 professors affiliated with Flemish universities and university colleges, makes a list of journals and publishers which, in their view, obeys the content related conditions. The 2012 approved journal list contained about 2800 non-WOS indexed titles and 125 publishers Those lists trigger the final inclusion of an institution's item into the VABB-SHW database.

A number of studies by ECOOM/University of Antwerp, manager/coordinator of the database, profiled the included publications (Engels, Ossenblok & Spruyt 2012). A study of Ossenblok, Engels & Sivertsen (2012) compares the VABB-SHW with the Nordic database Cristin (by which it was inspired) in terms of Web of Science coverage and language use. WOS coverage of SHW articles can vary substantially depending on discipline, with an average of 30-40%. However, contrary to the situation in Norway, this percentage is increasing for Flemish universities, potentially pointing to incentive effects caused by the absence of valorization of non-WOS publications in the period before 2008.

Whereas the citations of the Web of Science items are monitored by ECOOM, it is unknown what the citation impact is of non-WOS publications accepted for the VABB-

SHW. The threshold level being academic peer reviewed, the list of authorized titles contains journals with both national and international scope, in English or other languages. Therefore, the esteem of the authorized journals might vary, raising questions on how to measure the influence of the items included. The choice of publishers on the other hand is generally seen to be more selective, mainly covering international top scientific publishers. This study analyses the amount of WOS citations obtained by items accepted for the VABB-SHW database. The citations are divided by publication type.

Method

A set of 610 VABB-SHW approved articles was considered, authored by scientists of the Vrije Universiteit Brussel (VUB). The publication year is taken between 2002 and 2008. The citations were collected through a cited reference search in the online Web of Science. This study therefore considers citations of non-WOS, social sciences and humanities publications, by Web of Science included items. The methodology is in accordance with the extended citation analysis to non-sourced data proposed by Butler and Visser (2006). All citations are counted from the year of publication till 2013. The citation searches were performed mainly based on the publication title. This implies there might be citation variants undiscovered, making the presented data minimal estimations. Self citations, on the other hand, are included.

Results

Results are represented in Table 1. The first column gives the publication type, the second indicates the percentage of VABB-SHW publications being cited by WOS items, column 3 shows the CPP, the average citation per publication, whereas the last column represents the largest citation value obtained by a publication of that type.

Table 1: VABB-SHW citations in Web of Science

Publication type	# PUB	%-CIT	CPP	MCIT
Article in Journal	289	25%	1,63	68
Book as author	32	21,88%	1,34	12
Book as editor	41	21,95%	3,54	92
Chapter in book	248	26,61%	1,24	32

#PUB: number of publications per type in VABB-SHW (VUB) / **%-CIT:** percentage of publication receiving at least one citation / **CPP**: VABB-SHW number of citations per publication obtained with the online Web of Science Cited Reference Search/ **MCIT**: Maximum value of citation obtained by a publication

Figures indicate that the rate of cited non-source articles, book chapters or books by WOS-source items varies from 21% to 26%. On average each authored book or book chapter receives about 1,3 citations. For edited books this is substantially more, around 3 times the impact of an authored book. Almost a quarter of the VABB-SHW accepted articles are cited and the CPP equals 1,63. For book chapters, results are in line with the findings of Butler and Visser (2006) (taken over 5000 publications from 2 Australian Universities for the years 1997 and 1999) where about 30% of the items are cited. However, the Australian figures indicate a 65% citation for research monographs of commercial publishers contrary to just over 20% in the case of the VABB-SHW. The results are remarkable for two reasons: First, contrary to the Australian case, the VABB-SHW contains only social sciences and humanities items, where book contributions are expected to play a more prominent role and second because the AP's decision towards

publishers is considered to be more restrictive compared to journal titles.

Table 2: CPP VABB-SHW, AUSTRALIA, WOS: The discipline of Law

	Publication type	# Pub	# Cit	CPP
VABB	Articles in Journal	44	81	1,84
	Book as author	9	11	1,22
	Chapter in book	36	27	0,75
AU	Articles in Journal	584	142	0,24
	Book as author	36	150	4,17
	Chapter in book	228	91	0,40
WOS	Articles in Journal	24675	39314	1,59

VABB: figures for the Flemish Academic Database SHW / **AU**: Australia: Figures taken from Bulter and Visser (2006) / WOS: Average number publications period 2002-2008, citations to 2011: data sourced by Thomson Reuters Web of Knowledge (formerly referred to as ISI web of science) licence of ECOOM

Table 2 lights out the case of the discipline of Law, for the VABB-SHW and for the study of Bulter and Visser (2006). For journal articles, also the global numbers were taken for the matching WOS subcategory. Clearly not all figures are taken over the same time span, although the WOS window is close to that of the VABB-SHW.

For journal articles, the VABB-SHW CPP is in a close range with the WOS average and is higher than the CPP for Australia. The CPP for book chapters resembles the Australian rate more closely. In contrast, for books as author, the overall figures are confirmed, meaning that the number of citations per authored book are much smaller for the VABB-SWH [1,2 versus 4,2].

Conclusion

This paper analyses the citation impact of social sciences and humanities publications accepted for the Flemish Academic Bibliographic Database for Social Sciences and Humanities. Results are preliminary. As work in progress, only data for the Vrije Universiteit Brussel were considered, and elaboration is needed to publications of the other Flemish universities. However, some clear tendencies can be observed: For journal articles, where the list of approved titles is heterogeneous, a substantial percentage of items is referred to in international literature and the CPP is at about 1,6. For books and book chapters, values are slightly lower, but also a quarter of the items are cited and the CPP is higher than 1. The results add to the study of Ossenblok et al. (2012) that the VABB-SHW completes the WOS publications not only in terms of coverage, but also in terms of citation impact. Results are however not in full accordance with the study of Butler and Visser (2006) where mainly books as author stand out as the most important category in terms of CPP.

References

Butler, L. & Vissser M.S. (2006). Extending citation analysis to non-source items. *Scientometrics*, 66, 327-343.

Debackere, K. & Glänzel, W. (2004). Using a bibliometric approach to support research policy making: The case of the Flemish BOF-key. *Scientometrics,* 59, 253-276.

Engels, T.C.E., Ossenblok T.L.B. & Spruyt E.H.J. (2012). Changing publication patterns in the Social Sciences and Humanities, 2000-2009. *Scientometrics*, 93, 373-390.

Ossenblok, T., Engels, T. & Sivertsen, G. (2012). The representation of the social sciences and humanities in the Web of Science: a comparison of publication patterns and incentive structures in Flanders and Norway (2005-2009). *Research evaluation*, 21:4, 280-290.

COLLABORATION IN THE SOCIAL SCIENCES AND HUMANITIES: EDITED BOOKS IN ECONOMICS, HISTORY AND LINGUISTICS

Truyken L.B. Ossenblok[1] and Tim C.E. Engels[2]

[1] *Truyken.Ossenblok@ua.ac.be*
Centre for Research & Development Monitoring (ECOOM), University of Antwerp, Middelheimlaan 1, 2020 Antwerp (Belgium) (corresponding author)

[2] *Tim.Engels@ua.ac.be*
Department of Research Affairs and Centre for Research & Development Monitoring (ECOOM), University of Antwerp, Middelheimlaan 1, 2020 Antwerp (Belgium); Antwerp Maritime Academy, Noordkasteel-Oost 6, 2030 Antwerp, Belgium

Introduction

In order to quantify the degree of research collaboration, several collaborative measures based on mathematical computation of the number of co-authors and the co-occurrence of co-author pairs have been put forward. Scientific collaboration through co-authorship of journal articles is internationally on the rise, even though important differences between disciplines remain. Co-authorship is rarer in the social sciences and humanities (SSH) in particular. Moreover, the SSH disciplines are known to be underrepresented in databases such as the Web of Science and Scopus. As a result the collaboration network of social scientists and humanities scholars in particular cannot be faithfully reflected by the applying current collaboration measures to data extracted from these databases. Better measures of collaboration in the social sciences and humanities are needed (Sula, 2012)

By looking at collaboration as apparent from edited books, we want to make a bibliometric contribution to the better measurement of collaboration in the social sciences and humanities. The inclusion of books in the measurement of SSH collaboration is obviously relevant, as the publication output of SSH entails not only journal articles, but also a substantial proportion of books as well as chapters in edited books (Hicks, 2004; Nederhof, 2006; Engels, Ossenblok, & Spruyt, 2012)

Data and methods

For this research we used data from a Flemish full coverage database of peer reviewed publications in the SSH, the VABB-SHW (Engels et al., 2012). All publications 2000-2011 by SSH researchers affiliated with a Flemish university are indexed. However, only publications in journals and with publishers that have been the subject of peer review prior to publication, are eligible. All publications in the database are categorized in one or more disciplines, depending on the authors' departmental affiliation(s).

In our research an edited book is the traditional edited book where one or more editors coordinate the publication of a book to which a number of authors

contribute one or more chapters. The editors can also be one of the authors, as most of the time they (co)write the introduction and the conclusion and/or one of the chapters. We looked up details of the edited books 2000-2011 of three disciplines: Economics & Business, History and Linguistics. Per edited book the number of book chapters, authors per chapter and unique authors was harvested manually on the internet and in the university library. The results on 236 edited books, containing a total of 3.932 chapters, are presented in this poster. 34 books turned out not to be edited books (e.g. text editions written entirely by the editor), and thus were left out of this study, and 4 books were unavailable both online and through interlibrary document supply.

Results

In the three disciplines the average number of editors per edited book is larger than the average number of authors per book chapter within these edited books, i.e. 2,4 editors versus 1,1 authors for History; 2,7 versus 1,2 for Linguistics, and 3,1 versus 1,9 for Economics. The proportion of books edited by more than one editor is for all three disciplines well above 80% whereas the proportion of book chapters with more than one author are well below 53% (Economics). Thus we observe that across the three disciplines book editing is envisioned as a collaborative enterprise, typically undertaken by two to four people, perhaps because of the diversity of the tasks involved. This contrast with the co-authorship of the contributions, which only in Economics are typically the result of collaboration.

Table 12: Book rank by number of book chapters (BC) and by unique authors (AU)

book rank	Economics N= 55		History N = 49		Linguistics N= 132	
	#BC	#AU	#BC	#AU	#BC	#AU
1	63	134	87	94	50	55
2	52	98	34	34	49	54
3	47	74	30	29	44	48
4	44	67	22	23	43	47
5	43	66	22	22	42	44
6	38	63	22	21	40	41
7	35	60	22	21	35	40
8	22	46	22	21	33	36
9	20	37	21	21	30	34
10	20	34	21	20	30	33
11	19	31	20	20	28	29
12	19	30	19	19	26	29
13	17	26	19	19	26	29
14	17	25	17	18	26	28
15	16	25	17	18	26	28
16	16	25	17	18	25	27
17	16	25	16	17	24	26
18	16	24	15	17	24	25
19	15	23	15	16	24	25
20	14	23	14	16	24	25
21	14	23	14	16	23	24
22	14	22	14	16	22	24
23	13	22	13	15	22	23
…	…		…	…	…	…
N-4	9	12	9	8	8	8
N-3	9	12	8	8	8	7
N-2	8	12	8	8	7	5
N-1	8	9	7	7	6	4
N	8	6	7	7	3	2

Table 1 shows the number of book chapters and the number of unique authors per edited book, ranked independently for each discipline under study. The number of book chapters is an indication of the volume of the edited books, whereas the number of unique contributors illustrates the breadth of the network of the editor(s). To be able to measure both the volume and the breadth of the editors' network, we use a well-known measure, namely the H-index. The H-index is best known as an indicator of publication activity and citation impact, but can be defined more generally as a source-item relationship. In this study we defined the H-index as

the largest number n such that you have n edited books with n or more book chapters (H-BC), respectively unique authors (H-AU).

The H-BC and H-AU indicated in Table 1 demonstrate that within Economics 16 books have at least 16 book chapters and 22 books have at least 22 unique contributors. This difference between volume and number of contributors illustrates the collaborative publication pattern in edited Economics books. For History and Linguistics the H-BC and the H-AU are almost identical with 16 and 17 for History and 22 and 23 for Linguistics. This is in line with the large percentage of book chapters written by a single author. However, edited books with one chapter written by a large number of unique authors combined with a large number of chapters by one and the same author, although very unlikely in a humanities discipline, can give the same results. The data on the tail (book ranks N to N-4) of the distribution illustrate that less than 5% of edited books contain contributions by 8 or less authors. The full data series illustrate that in History and Linguistics the majority of the edited books contain contributions by respectively 15 to 24 and 17 to 28 authors, whereas in Economics there are 23 to 40 contributors in a majority of cases. Thus, in Economics book chapters involve more collaboration than in History and Linguistics. In sum, the inconsistencies of the H-index notwithstanding, this poster illustrates that the concept of an H-index for edited books is fruitful and needs further exploration.

Conclusion

Book chapters and (edited) books are an important scientific publication medium within the SSH. Using the H-index on book chapters and unique contributors to an edited volume, we indicated the difference between three disciplines in both volume of an edited book and in breadth of the editors network. Additional data from other SSH disciplines will allow us to provide a broader perspective and to draw more general conclusions. The results furthermore underline the difference in nature between co-authoring and co-editing and hence the need to take into account the (special) relation between editors and chapter authors when studying collaboration patterns. A new measure for calculating collaboration will be investigated in future research.

References

Engels, T. C. E., Ossenblok, T. L. B., & Spruyt, E. H. J. (2012). Changing publication patterns in the social sciences and humanities, 2000-2009. *Scientometrics*, 93, 373-390.

Hicks, D. (2004). The four literatures of social science. In H.F.Moed, W. Glänzel, & U. Schmoch (Eds.), *Handbook of quantitative Science and Technology Research: The use of publication and patent statistics in studies of S&T systems* (pp. 473-496). Dordrecht: Kluwer Academic.

Nederhof, A. J. (2006). Bibliometric monitoring of research performance in the Social Sciences and the Humanities: A review. *Scientometrics*, 66, 81-100.

Sula, C. A. (2012). Visualizing Social Connections in the Humanities: Beyond Bibliometrics. *Bulletin of the American Society for Information Science and Technology*, 38, 31-35.

THE COLLECTIVE CONSEQUENCES OF SCIENTIFIC FRAUD: AN ANALYSIS OF BIOMEDICAL RESEARCH

Philippe Mongeon[1] and Vincent Larivière[1]

[1] *philippe.mongeon@umontreal.ca; vincent.lariviere@umontreal.ca*
École de bibliothéconomie et des sciences de l'information, Université de Montréal, C.P. 6128, Succ. Centre-Ville, Montréal, QC. H3C 3J7 Canada

Introduction

The issue of scientific misconduct has been largely discussed in the scientific community over the last few years. Major fraud cases shocked the scientific community (e.g., Woo-Suk Wang, Eric Poehlman, Diederik Stapel), and the number of papers retracted each year from the biomedical literature increased tremendously. Scientific misconduct, which includes fabrication, falsification and plagiarism (Merton 1973; Steneck 2006), accounts for about 53% of these retractions (Ferric 2012). Such behaviour has consequences for the scientific community (waste of resources, misleading of further research, loss of the public's trust), for the public (waste of public funds, ill advised decisions potentially affecting the health of individuals), and of course for the person responsible of the misconduct (Steneck 2006). Past studies have looked at the number of citations received by retracted papers after they were retracted. While some have shown an important decline in the number of post-retraction citations (Furman et al (2012), others only observed a small decline (Budd et al. 1998; Pfeifer and Snodgrass 1990) or even no significant decline at all (Neale et al. 2007; Neale et al. 2010). Similarly, Neale et al. (2010) have shown that citing authors were not aware of the retraction.

It is assumed that co-authors of retracted papers are affected by their colleague's misconduct (Bonetta 2006), but no study has yet attempted to measure the consequence of such retractions on their ulterior research careers. The purpose of this study is, thus, to provide empirical evidence of these consequences, by measuring the pre and post retraction productivity and scientific impact of collaborators of retracted papers. This study is timely, considering that the increase in the number of retractions, combined with an increase in number of authors per paper (Larivière et al. 2006), allow us to expect that the number of "innocent" co-authors affected by the misconduct of their collaborators will continue to increase in the next few years.

Methods

We used PubMed to find all retraction notices between 1996 and 2006 as well as the corresponding retracted paper. The retracted papers were then found on the Web of Science (474), and provided us with a list of 1,920 distinct authors. Retractions in biomedical and clinical research (443 retractions and 1,818 authors) were then categorized by reason for retraction using data from Azoulay et al. (2012) or by reading the retraction notice. The reason for retraction was data fabrication or

falsification in 155 cases (35%) and plagiarism in 26 cases (5.9%) for a total of 181 (40.9%) cases of misconduct affecting 633 authors (34.8%). In order to focus on the impact on the "innocent" collaborators, we used the data from Azoulay et al. (2012) and the retraction notices to identify, for each retracted publication, the author(s) that were officially found responsible for the misconduct. 67 authors (10.6%) were identified as responsible for 144 (79.6%) of the retracted publication for misconduct and were excluded of our analysis. We then searched the WoS for all citable publications by all remaining 1,751 authors within a range of 5 years before and after the retraction, providing us with a sample of 42,011 distinct publications, covering 1991-2011.

To measure the impact of the retraction on researchers' ulterior careers, we calculated for each co-author: 1) the average number of citations received by the co-author's paper normalized by year and discipline of journal; 2) the number of published papers and 3) the number of authors per paper. We measured those 3 indicators for 5 years before the retraction and for 5 years after. For authors with retractions on different years, data was collected for 5 years before the first retraction and 5 years after the last. This reduced our final sample to 37,436 articles and 1,719 authors (629 for cases of misconduct).

Results

In cases of misconduct, our results show that the number of researchers that score above average for citations received (i.e. lower than one) decreases by about 5% after the retraction (See Figure 1) while the number of researchers that score under average increases. There is no such decrease in cases of retractions for other reasons, where we instead observed a small increase in the number of researchers scoring above average.

Our results (Figure 2) also show that the number of articles per author is declining similarly for both cases of misconduct and other reasons of retraction.

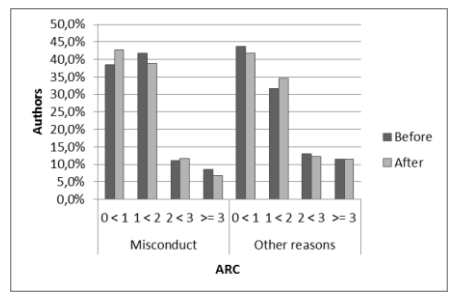


Figure 1. Average of relative citations (ARC) of co-authors' publications 5 years before and after retractions

In both cases, we found that a significant proportion of authors (about 25%) did not publish in the 5 years following the retraction. The strongest decrease is found among the researchers who have between 1 and 10 publications, dropping from 56,7% to 40,2% in the case of misconduct, and from 59,4% to 43,7% in other cases. This suggests that the retraction of a publication can have a decisive effect on the desire (or ability) of individuals to pursue a scientific career and that this effect is potentially stronger amongst researchers who already publish less than others.

As for the number of authors per paper, we see no significant trend before and after retractions, both in the case of misconduct and other reasons. This suggests that retractions have no clear impact on collaborative practices of co-authors.

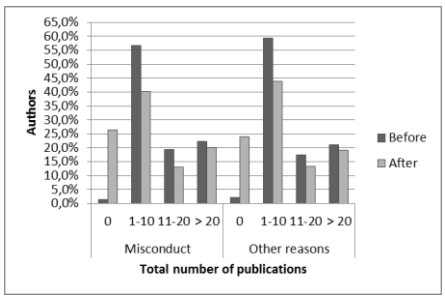


Figure 2. Total number of publications by authors 5 years before and after retractions

Discussion

Our findings validate the assumption that co-authors are also experiencing the consequences of the misconduct of their colleagues. This is consistent (and perhaps related) with the findings of Azoulay et al. (2012), indicating that whole research fields are likely to feel the impact of misconduct cases (e.g., decrease of new entry and funding). Our results provide yet another evidence of threat that misconduct poses for the scientific community, which calls for measures to be taken in order to reduce its prevalence not only in the biomedical field, but in the entire scientific community.

References

Azoulay, P., Furman, J. L., Krieger, J. L., & Murray, F. E. 2012. Retractions. *NBER working paper series*, 18449.

Bonetta, Laura. 2006. The aftermath of scientific fraud. *Cell*, 124 (5), 873-5.

Budd, JM, ME Sievert, & TR Schultz. 1998. Phenomena of retraction - reasons for retraction and citations to the publications. *Journal of the American Medical Association*, 280 (3), 296-7.

Ferric C. Fang, R. Grant Steen, & Arturo Casadevall. 2012. Misconduct accounts for the majority of retracted scientific publications. *PNAS*, 109 (42), 17028-17033.

Furman, Jeffrey L., Kyle Jensen, & Fiona Murray. 2012. Governing knowledge in the scientific community: Exploring the role of retractions in biomedicine. *Research Policy*, 41 (2), 276-90.

Larivière, V., Gingras, Y. & Archambault, É. 2006. Canadian collaboration networks: A comparative analysis of the natural sciences, social sciences and the humanities. *Scientometrics*, 68 (3), 519-533.

Merton, Robert K. & Norman W. Storer. 1973. *The sociology of science: Theoretical and empirical investigations*. Chicago: University of Chicago Press.

Neale, Anne et al. 2007. Correction and use of biomedical literature affected by scientific misconduct. *Science and Engineering Ethics*, 13 (1), 5-24.

Neale, Anne Victoria, Rhonda K. Dailey, & Judith Abrams. 2010. Analysis of citations to biomedical articles affected by scientific misconduct. *Science and Engineering Ethics,* 16 (2), 251-61.

Pfeifer, Mark P. & Gwendolyn L. Snodgrass. 1990. The continued use of retracted, invalid scientific literature. *The Journal of the American Medical Association,* 263 (10),1420.

Steneck, Nicholas. 2006. Fostering integrity in research: Definitions, current knowledge, and future directions. *Science and Engineering Ethics,* 12 (1), 53-74.

COMPARING NATIONAL DISCIPLINARY STRUCTURES: A QUANTITATIVE APPROACH

Irene Bongioanni[1], Cinzia Daraio[2] and Giancarlo Ruocco[1]

[1] *irene.bongioanni@uniroma1.it, giancarlo.ruocco@roma1.infn.it*
Sapienza University, Department of Physics, Piazzale Aldo Moro 5, 00185 Rome (Italy)

[2] *daraio@dis.uniroma1.it*
Corresponding author
Sapienza University, Department of Computer, Control and Management Engineering Antonio Ruberti, Via Ariosto 25, 00185 Rome (Italy)

Abstract

In this paper we propose a quantitative measure to compare the diversity of disciplinary profiles of countries. It is based on the theory of complex systems in physics, and in particular on the spin-glasses literature (Parisi, 1983; Mezard et al 1984a,b,; Fisher and Hertz, 1991). Our Diversity of Disciplinary Profiles (DDP) measure ranges from -1 to 1. The investigation on the distribution of the DDP is particularly interesting: if it shows a peak on 1 it means that there is a convergence of all analysed units towards a unique profile in their scientific specialization; if it presents two peaks, it points to a kind of 'broken symmetry' situation in which two different configurations or patterns of scientific specialization emerge. Finally, a broad distribution of the DDP index indicates fully independence of the scientific profiles of the different countries.

We analyse the number of publications (integer count) of European countries in the 27 Scopus subject categories over 1996-2011. We compare the disciplinary profiles of European countries i) among them; ii) with respect to the European standard; and iii) to the World reference. The distributions of the resulting DDP show that there is a convergence towards a unique European disciplinary profile, confirming a trend of globalization of science in Europe.

Introduction

The disciplinary structure of the scientific production of countries has been much studied in the literature. Several studies have analysed national publication profiles. National publication profiles indeed show interesting features about a country's research system and its national scientific policy. Recent works include Glanzel et al. (2008), Almeida et al. (2009), and Zhang et al. (2011).

A commonly used approach is based on the study of publication profiles by discipline. Within this framework, the world's scientific output is divided into major scientific fields, and the relative contribution of each country with respect to each field is illustrated on a radar chart. The publication profile of a national research system is then measured by the Relative Specialization Index which indicates whether a country has a relatively lower or higher share in world publications in a given discipline than in its overall share of world total

publications. Beside, several measures of similarities or diversities (dissimilarities) over given categories have been proposed, including the Pratt index, the Shannon entropy, the Stirling's diversity and the Stirling-Rao diversity measure (see for a recent systematization Stirling, 2007).
On the contrary, much less explored in the literature is the quantitative evaluation of scientific production profiles and the investigation of the distribution of their similarity or diversity measures.
The main objective of this paper is to propose a quantitative measure to assess the similarity or diversity of countries disciplinary specialization and investigate its resulting distribution.

Method

The main variables analysed here are the $P^a(i)$, i.e. the share of articles (integer counting) published in a subject category *i* for a given country *a* over the sum of publications in 1996-2011
At this purpose, we standardize their values as follows:

$$\sigma^a(i) = \frac{P^a(i) - \langle P^a(i) \rangle}{\sqrt{\langle P^a(i)^2 \rangle - \langle P^a(i) \rangle^2}}$$

where $\langle \circ \rangle$ stands for average of °.
These $\sigma^a(i)$ have the following properties:

$$\langle \sigma^a \rangle = 0 \quad \text{and} \quad \langle [\sigma]^a)^2 \rangle = 1$$

Then, the measure of diversity of profiles between research systems *a* and *b*, named as *Diversity of Disciplinary Profile index*, *DDP,* called O(*a,b*), can be calculated as follows:

$$O(a,b) = \frac{1}{N}\sum_{i=1}^{N} \sigma^a(i)\sigma^b(i)$$

where *i* denotes the subject category and *N* is the total number of subject categories, in our case 27.
Our DDP measure of similarity or dissimilarity of profiles ranges from -1 meaning precisely opposite profile, to 1 meaning precisely the same profile, with 0 representing independence, and intermediate values indicating in-between level of similarity or dissimilarity.
Moreover, the examination of the distribution of the overlaps is particularly interesting. If it shows a peak on 1 it means that there is a *convergence* of all analysed units towards a *unique* profile in their scientific specialization; on the contrary if it presents two peaks, it points to a kind of 'broken symmetry' situation in which two different configurations or patterns of scientific specialization emerge.
In addition, the overlap can be calculated with respect to *another country*, or with respect to an *average* or standard value, or with respect to a *given distribution*.

Data

Data come from Scopus database and refer to the scientific production of 27 European countries in the 27 Scopus subject categories (disciplines) from 1996 to 2011, including the total world scientific production by discipline as a reference. The available variables include: number of articles (including articles, reviews and conference proceedings papers) in integer and fractional counts; total number of citations on a 4 year window, relative citation impact, number of articles in top 10% of most highly cited articles in a discipline, number of internationally co-authored papers, number of nationally co-authored papers and number of single authored papers.

Table 1. Diversity of Disciplinary Profile Indices among each country and the European and World standard. In the bottom of the table some descriptive statistics are reported.

Country	*standard Europe*	*standard World*
AUT	0.994	0.954
BEL	0.995	0.960
BGR	0.853	0.805
CYP	0.691	0.767
CZE	0.953	0.908
DEU	0.989	0.948
DNK	0.952	0.891
ESP	0.984	0.944
EST	0.840	0.816
FIN	0.978	0.968
FRA	0.990	0.956
GBR	0.970	0.952
GRC	0.968	0.973
HUN	0.944	0.882
IRL	0.983	0.974
ITA	0.993	0.952
LTU	0.738	0.773
LUX	0.863	0.881
LVA	0.670	0.718
MLT	0.851	0.892
NLD	0.967	0.934
POL	0.931	0.891
PRT	0.878	0.892
ROU	0.658	0.720
SVK	0.890	0.838
SVN	0.853	0.901
SWE	0.982	0.942
Min	0.658	0.718
Max	0.995	0.974
Mean	0.902	0.890
Std. Dev.	0.105	0.077

Results

By applying the methodology described above, we compare the disciplinary profiles of European countries 1) between them, 2)with respect to the European standard and 3) with respect to the World reference. We consider the $P^a(i)$, i.e. the shares of articles (integer counting) published in a subject category *i* for a given country *a*.

In Table 1 the detailed values of DDP between each country and the European and World standard are reported. Direct comparisons of the "distances" of each country from the standards (Europe and World) are made possible thanks to the descriptive statistics reported at the bottom of Table 1, including minimum and maximum, as well as mean values and standard deviations.

Figure 1 shows the distribution of DDP indices calculated on the indicator PUB (number of publications in integer counting) among European countries. We can observe that this distribution presents a well-defined peak near 1, meaning that European countries have substantially equivalent specialization profiles. Figure 2 reports the same distribution for the indicator PUBf (number of papers in fractional counting).

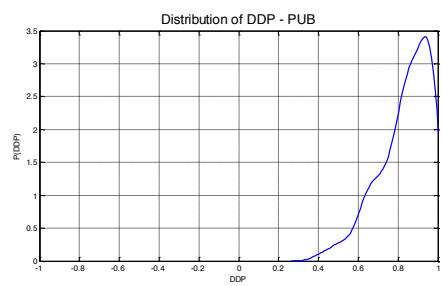


Figure 1. Nonparametric kernel distribution of the DDP indices among European countries. Stock of Publications (1996-2011) - integer counting.

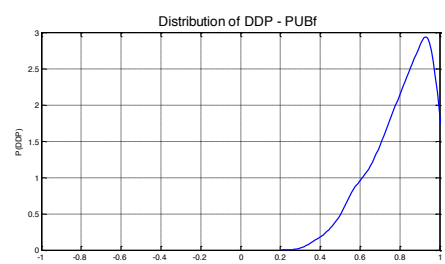


Figure 2. Nonparametric kernel distribution of the DDP indices among European countries. Stock of Publications (1996-2011) - fractional counting.

Figure 3 illustrates the evolution of the average DDP indices of European countries by year, from 1996 to 2011. It shows that over time there has been a convergence of European countries over a unique scientific specialization profile, confirming a trend of globalization of science in Europe.

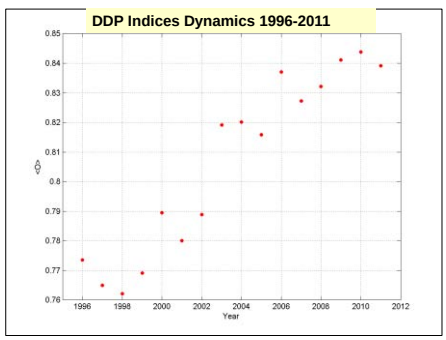


Figure 3. Dynamics of DDP indices over 1996-2011 – PUB integer counting.

Further developments

Having showed the usefulness of the proposed approach, there are several extensions that will be pursued in further research. Namely:

- extending the analysis to all world countries;

-extending the investigation at different level of analysis (e.g. at regional level);

-investigating the dynamics of the disciplinary profiles over time,

-analysing the behaviour of other scientific production indicators, not analysed in this paper, such as citations, relative citation impact, number of articles in top 10% of most highly cited articles, number of internationally co-authored papers, number of nationally co-authored papers and number of single authored papers.

Acknowledgments

Data have been provided by Elsevier within the Elsevier Bibliometric Research Project "Assessing the Scientific Performance of Regions and Countries at Disciplinary level by means of Robust Nonparametric Methods: new indicators to measure regional and national Scientific Competitiveness".

Selected References

Almeida J.A.S., Pais A.A.C.C., Formosinho S.J. (2009), Science indicators and science patterns in Europe, *Journal of Informetrics*, 3, 134-142.

Fisher, K. H., Hertz, J. A. (1991). *Spin Glasses*. Cambridge University Press.

Glanzel W. (2000), Science in Scandinavia: A bibliometric approach, *Scientometrics*, 48(2), 121-150.

Glanzel W., Debackere K., Meyer M. (2008), 'Triad' or 'tetrad'? On global changes in a dynamic world, *Scientometrics*, 74 (1), 71-88.

Mezard M., Parisi G. Sourlas N. Toulouse G., Virasoro M. (1984a) Nature of the Spin-Glass Phase, *Physical Review Letters*, 52 (13), 1156-1159.

Mezard M., Parisi G. Sourlas N. Toulouse G., Virasoro M. (1984b) Replica symmetry breaking and the nature of the Spin-Glass Phase, *Journal de Physique*, 45 (13), 843-854.

Parisi, G. (1983), Order parameter for Spin-Glasses, *Physical Review Letters*, 50 (24), 1946-1948.

Stirling, A. (2007). A general framework for analysing diversity in science, technology and society. *Journal of the Royal Society Interface*, 4(15), 707–719.

van Raan A.F.J. (2004), Measuring science, in H.F. Moed, W. Glanzel and U. Schmoch

(edited by), Handbook of Quantitative Science and Technology Research, Kluwer Academic

Publishers, pp. 19-50.
Zhang, L., Rousseau R, Glanzel W. (2011), Document-type country profiles, *Journal of the American Society for Information Science and Technology*, 62(7),1403-1411.

COMPREHENSIVENESS AND ACCURACY OF DOCUMENT TYPES: COMPARISON IN WEB OF SCIENCE AND SCOPUS AGAINST PUBLISHER'S DEFINITION

Kazuhiro Hayashi[1] and Nobuko Miyairi[2]

[1] *khayashi@nistep.go.jp*
National Institute of Science and Technology Policy, 3-2-2 Kasumigaseki, Chiyoda-ku, Tokyo 100-0013 (Japan)

[2] *n.miyairi@nature.com*
Nature Publishing Group, Chiyoda Bldg., 2-37 Ichigaya Tamachi, Shinjuku-ku, Tokyo 162-0843 (Japan)

Background

The document type is an important bibliographic element and used in many bibliometric studies to segment source data into a meaningful set by different nature of scholarly communication. Articles, often regarded as most substantial body of research, constitute the primary source for creating bibliometric indicators. Other document types are often treated differently; some are disregarded when calculating indicators.

Past studies

Since early days of bibliometric research, document types have been used as a facet of analysis (Braun, Glänzel & Schubert, 1989). Science Citation Index and its siblings by Institute for Scientific Information have been traditionally the main tool of such analyses; however, as more services became available in recent years, there are studies reporting the coverage and accuracy of different systems (Bar-Ilan, Levene & Lin, 2007; Jacsó, 2009; Michels & Schmoch, 2012). Interpretation of document types can vary by subject domain (Sigogneau, 2000; Harzing, 2013). The assignment of document types is usually controlled by the database vendor and it influences the calculation of bibliometric indicators and even triggers controversies (Rossner, Van Epps, & Hill, 2007).

Purpose of the study

The purpose of this study is to examine how document types are treated in different services, and consider its implications for bibliometric research in general. Our analysis is descriptive and not intended to draw any statistical inference. This is a study in progress and we report our initial observations below.

Data collection

We collected database records for 18 journals of Nature Publishing Group (NPG) that were published in the years 2009-2011 from two services: Thomson Reuters' Web of Science (WoS) and Elsevier's SciVerse Scopus (Scopus). We also used Nature Publishing Index (NPI) content management system, which indexes NPG's primary research journals including those 18 titles, in

order to cross-check the numbers obtained from WoS and Scopus.

We used the full journal title for searching in each database, then broke down the results by publication year and downloaded only the records tagged as "2009", "2010", or "2011". Note at this point we did not discriminate any document type.

Table 1. Number of items indexed in WoS and Scopus for NPG journals published in 2009-2011

Journal	WoS	Scopus
Nature	7712	7444
Nature Biotechnology	1059	1133
Nature Cell Biology	684	719
Nature Chemical Biology	599	617
Nature Chemistry	695	704
Nature Climate Change	201	0
Nature Communications	602	607
Nature Genetics	817	822
Nature Geoscience	842	799
Nature Immunology	629	648
Nature Materials	881	713
Nature Medicine	1498	1493
Nature Methods	905	901
Nature Nanotechnology	641	657
Nature Neuroscience	869	888
Nature Photonics	745	672
Nature Physics	817	789
Nature Structural & Molecular Biology	742	717
Total	**20938**	**20323**

Initial findings

WoS vs. Scopus

The number of items indexed from each of 18 journals often does not agree between WoS and Scopus (Table 1). Scopus tends to have more items in life science journals, while WoS tends to have a greater number in physical science journals.

Taking *Nature* as an example, we examined how the numbers compare by document type (Table 2). Between WoS and Scopus, "Article", "Letter" and "Review" seemed reasonably comparable.

Table 2. Number of *Nature* records indexed in WoS and Scopus for publication years 2009-2011

Database	WoS			Scopus		
Year	2009	2010	2011	2009	2010	2011
Total	2544	2577	2591	2411	2492	2541
Article	800	825	804	881	899	1076
Letter	250	268	283	217	231	258
Review	66	37	37	102	49	74
Editorial (Material)	780	760	818	195	196	169
Correction (Erratum)	78	68	79	60	73	68
News Item	381	448	401			
Book Review	169	150	147			
Biographical Item	19	21	22			
Reprint	1					
Note				511	633	603
Short Survey				442	409	277
Article in Press				3	2	15
Conference Paper						1

Table 3. Comparison of three document types published in *Nature* in 2009-2011

Year / Doc Type	NPI	WoS	Scopus
2009 / Article	117	800	881
2009 / Letter	666	250	217
2009 / Review	20	66	102
2010 / Article	145	825	899
2010 / Letter	669	268	231
2010 / Review	13	37	49
2011 / Article	136	804	1076
2011 / Letter	658	283	258
2011 / Review	16	37	74

WoS and Scopus vs. NPI

The numbers obtained from each database show considerable disagreement when compared with NPI (Table 3). There are more "Article" and "Review", and fewer "Letter" than publisher's definitions.

These 3 document types were matched by DOI for further investigation, and results were as follows:

- Among 398 items labelled as "Article" in NPI, WoS classifies 397 as "Article" and 1 as "Review". In Scopus, 394 of them were found as "Article" and 4 as "Review".
- All 1993 items identified as "Letter" in NPI were treated as "Article" in WoS. In Scopus, 1979 appeared as "Article", 10 as "Article in Press", 13 as "Letter", and 1 as "Review".
- All 49 "Review" items in NPI were classified as "Review" in both WoS and Scopus.

In Scopus, there were total of 111 DOIs duplicating in more than one record. Among 20 items indexed as "Article in Press" in Scopus, 12 had duplications labelled as "Article". There were no such duplications found in WoS.

Implications for further research

Both databases classify items in the "Letter" section of *Nature* as "Article", although results in Scopus are somewhat mixed. This seems reasonable since those items may appear shorter in length but often communicate primary research findings. Items defined as "Articles" and "Review" by NPG mostly classified so in both database; however more granular and systematic comparisons would be required to understand a few exceptions.

On the other hand, there are 40 more "Article" and 801 "Letter" in WoS, and 483 more "Article" and 693 "Letter" in Scopus, which belong neither of those sections of *Nature*.

The composition of these document types is of our particular interest for broader implications; for example, if non-primary research publications are classified as "Article", it may result as a diluting effect in the calculation of certain indicators. We plan to extend our analysis to the rest of document types and other journals retrieved in our search.

References

Bar-Ilan J., Levene M., & Lin A. (2007). Some measures for comparing citation databases. *Journal of Informetrics*, *1*(1), 26-34.

Braun T., Glänzel W., & Schubert A. (1989). Some data on the distribution of journal publication types in the science. *Scientometrics*, *15*(5-6), 325-330.

Elsevier B. V. SciVerse Scopus. Accessed January 4, 2013 from: http://www.scopus.com/

Harzing, A.W. (2013). Document categories in the ISI Web of Knowledge: Misunderstanding the Social Sciences? *Scientometrics*, 93(1), 23-34.

Jacsó P. (2009). Errors of omission and their implications for computing scientometric measures in evaluating the publishing productivity and impact of countries. *Online Information Review*, *33*(2), 376-385.

Michels, C., & Schmoch, U. (2012). The growth of science and database coverage. *Scientometrics*, *93*(3), 831–846.

Nature Publishing Group. *Nature Publishing Index.* Public access available from: http://www.natureasia.com/en/publishing-index/

Rossner, M., Van Epps, H., & Hill, E. (2007). Show me the data. *The Journal of Cell Biology*, *179*(6), 1091–2.

Sigogneau A. (2000) An Analysis of Document Types Published in Journals Related to Physics: Proceeding Papers Recorded in the Science Citation Index Database. *Scientometrics*, *47*(3), 589-604.

Thomson Reuters. *Web of Science*. Accessed January 4, 2013 from: http://isiknowledge.com/

CONTRIBUTION OF BRAZILIAN SCIENTIFIC PRODUCTION TO MAINSTREAM SCIENCE IN THE FIELD OF MATHEMATICS: A SCIENTOMETRICS ANALYSIS (2002-2011)

Renata Cristina Gutierres Castanha[1] and Maria Cláudia Cabrini Grácio[2]

[1] *regutierres@gmail.com*
UNESP – Univ Estadual Paulista, 737 Hygino Muzzi Filho Avenue, 17525-900 Marília (Brazil)

[2] *cabrini@marilia.unesp.br*
UNESP – Univ Estadual Paulista, 737 Hygino Muzzi Filho Avenue, 17525-900 Marília (Brazil)

Introduction

Mathematics is an area of research with great international collaboration due to its peculiarities, such as language understood worldwide and little restriction in relation to material resources for carrying out researches, which are mostly theoretical (Dang & Zhang, 2003). Luksch & Behrens (2011) presented a bibliometric study of Mathematics in the period 1868- 2008 worldwide.

In Brazil, the scientific production in Mathematics takes the 16th position in production rankings, as seen in SCImago Country Rank, in period 1996-2011. Therefore, the need to assess the Brazilian scientific production in Mathematics is highlighted, given the lack of studies in the area and the importance of mathematical studies that support the framework of different areas, contributing to the development of science as a whole.

In this context, bibliometric indicators of production, citation and scientific collaboration contribute to identify the current scenario of Brazilian scientific production in Mathematics indexed in mainstream science. Production indicators contribute to evidence, among other characteristics, researchers, institutions, journals and countries in a scientific community, enabling the identification of their prolific producers. Papers have become the most popular basic unit for bibliometric analysis once they present original research results, are submitted by a review system based on evaluation rules, and compose broad access literature. The citation indicators show impact and visibility of an author, journal or country within their community. Collaboration join efforts to provide better researching conditions for the group involved, offering support, exchange, information and resource sharing. For studies on collaborative analysis at macro level, such as those among countries, scientific collaboration is well portrayed by co-authorships of published papers (Glänzel, 2003; Glänzel & Schubert, 2004).

In light of the presented issues, this research aims to conduct a diachronic study of the scientific production in the field of Mathematics with the presence of Brazilian researchers, based on Scopus from 2002 to 2011, identifying its impact on the international scientific

community. Moreover, the study highlights the major journals in Mathematics where these publications were disseminated, the most productive Brazilian institutions in this set of scientific publications and the main Brazilian collaborator countries in Mathematics studies.

Methodological procedures

From the search in Scopus, in the field of Mathematics, 12,240 articles with at least one author from Brazil were retrieved, published from 2002 to 2011. This study considered the prolific institutions those are responsible for more than 1.5% of published articles in Mathematics in the period. This threshold (1.5%) was also used to analyze main journals and collaborator countries.

For each year, the annual growth rate of Brazilian scientific production in the area was also calculated. Also, the total number of citations received by articles with Brazilian authors was obtained, per year. From this data, it was possible to calculate the citations per article per year.

Table 1. Total of articles, annual growth rate, production percentage and citations.

Year	*# articles*	*annual growth*	*% production*	*Cit per article*
2002	843	-	2.2%	11.6
2003	939	11.4%	2.0%	8.6
2004	1271	35.4%	2.3%	6.8
2005	929	-26.9%	2.0%	8.3
2006	1060	14.1%	2.0%	7.5
2007	1121	5.8%	1.9%	7.1
2008	1336	19.2%	1.9%	5.7
2009	1590	19.0%	1.8%	3.8
2010	1604	0.9%	1.9%	2.4
2011	1727	7.7%	1.9%	1.0
Total	12420	104.9%	2.0%	5.6

Presentation and analysis of data

Table 1 shows the diachronic distribution of production and citation indicators of the Brazilian articles published in Mathematics (2002-2011).

Except in 2005, the annual growth rate is always positive, more significantly in 2004, 2008 and 2009. In the accumulated period, Brazilian scientific production in the area more than doubled presenting accumulated growth rate of 104.9%.

Despite this growth, Brazilian participation in the area, as measured by the percentage of the production, remained around 2% throughout the period, slightly above this percentage in the first years and below that at the end of the period.

Regarding citations, a decreasing trend was observed: the most cited articles were published in 2002 and 2003, whereas in recent years the articles received fewer than 5 citations on average.

Brazilian scientific production has been disseminated by 6 major journals: Physical Review E-Statistical, Nonlinear, and Soft Matter Physics (7.5%); Physica A: Statistical Mechanics and Its Applications (5.1%); Physical Review D: Particles, Fields, Gravitation and Cosmology (4.6%); Lecture Notes in Computer Science (4.1%); Journal of Physics A: Mathematical and Theoretical (3.2%); Journal of Mathematical: Analysis and Applications (1.6%). None of these journals is Brazilian, three are European and three are North American. Four of these journals are in the first quartile of the journals ranking according to SCImagoJR, indicating that the Brazilian production has been inserted into high visibility channels.

It is noteworthy that out of the 12,240 analyzed articles, 4,764 (38.9%) were published by Brazilian authorship with at least one foreign author, thus indicating that a significant part of Brazilian production in the area has been developed internationally. Among

the 85 collaborator countries, 11 main countries were: United States (10.1%), France (6.6%), Germany (3.9%), United Kingdom (3.7%), Spain (3.6%), Italy (3.3%), Chile (2.4%), Russian Federation (2.4%), Canada (2.3%), Portugal (2.1%) and Argentina (1.8%).

Table 2 presents the 21 most productive Brazilian institutions in the area. All institutions most productive have graduate programs and that the graduate programs from USP, UNICAMP, UFRJ, IMPA and UFMG have equivalent performance with international centers of excellence.

Most productive institutions are public: 15 federal institutions and five state universities. Among state universities, three São Paulo public universities are highlighted, with sponsorship by São Paulo Research Foundation (FAPESP). This foundation has achieved international reputation and made agreements with research councils in several countries, including Holland, France, USA, Canada, Germany and the UK (Gibney, 2012). University of São Paulo, the largest producer of scientific papers in Mathematics, is on the top of Latin-American rank in Times Higher Education World University Rankings (Gibney, 2012).

Table 2. Most productive Brazilian Institutions.

Institution (abbreviation)	# articles	%
Univ. de São Paulo (USP)	2759	22.5%
Univ. Est. Campinas (UNICAMP)	1329	10.9%
Univ. Fed. Rio de Janeiro (UFRJ)	1182	9.7%
Univ. Est. Paulista (UNESP)	714	5.8%
Inst. Nac. de Mat. Pura e Apl. (IMPA)	631	5.2%
Univ. Fed. de Minas Gerais (UFMG)	613	5.0%
Univ. Federal de Pernambuco (UFPE)	534	4.4%
Univ. Fed. Rio Grande Sul (UFRGS)	500	4.1%
Univ. Federal Fluminense (UFF)	488	4.0%
Univ. de Brasília (UnB)	482	3.9%
Centro Br. de Pesq. Físicas (CBPF)	404	3.1%
Univ. Federal do Ceara (UFC)	375	3.1%
Pont. Univ. Cat. R. Janeiro (PUC-Rio)	342	2.8%
Univ. Fed. de São Carlos (UFSCAR)	320	2.6%
Univ. Federal da Paraíba (UFPB)	303	2.5%
Univ. Fed. Santa Catarina (UFSC)	300	2.5%
Univ. do Est. Rio de Janeiro (UERJ)	289	2.4%
Univ. Estadual de Maringá (UEM)	288	2.4%
Univ. Federal do Paraná (UFPR)	275	2.2%
Lab. Nac. Comp. Cientifica (LNCC)	244	2.0%
Univ. Fed. Rio Grande Norte (UFRN)	197	1.6%

Final considerations

Brazilian scientific production presented a positive annual growth. Most of major journals that disseminate Brazilian scientific production are in the first quartile of the Mathematics journal ranking. The most productive institutions have graduate programs and are public universities. A significant part of Brazilian production has been developed in international scientific collaboration.

References

DANG, Y. & ZHANG, W. (2003). Internationalization of mathematical research. *Scientometrics*, 58 (3), 559-570.

GLÄNZEL, W. (2003). *Bibliometrics as a research field*: a course on theory and application of bibliometric indicators. Retrieved from: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.97.5311&rep=rep1&type=pdff .

GLÄNZEL, W. & SCHUBERT, A. (2004). Analyzing scientific

networks through co-authorship. In Moed et al (eds.), *Handbook of Quantitative Science and Technology Research* (pp. 257-276). Netherlands: Kluwer Publishers.

GIBNEY, E. (2012). Boom times and golden goals. *Times Higher Education.* Retrieved December 28, 2012 from: http://www.timeshighereducation.co.uk/story.asp?storycode=422083.

LUKSCH, P. & BEHRENS, H. (2011). Mathematics 1868-2008: a bibliometric analysis. *Scientometrics*, 86 (1), 179-194.

CO-OCCURRENCE BETWEEN AUTHORS' AFFILIATION AND JOURNAL: ANALYSIS BASED ON 2-MODE NETWORK

Wei Ruibin[1, 2], Wu Yishan[1], Tian Dafang[2]

[1]*rbwxy@126.com, Wuyishan@istic.ac.cn*
Institute of Scientific and Technical Information of China, Beijing, (China)

[2]*tdf7011@126.com*
Anhui University of Finance and Economics, Bengbu,Anui (China)

Introduction

Co-words network and co-citation network may reveal the relationship among words or papers according to their co-occurrence. These networks usually appear as 1-mode network. This paper aims to reveal the relationship between academic institutions and their publications based on 2-mode network. Such network can bring out more abundant information than 1-mode network. Through the institutes-journals 2-mode network, we can find the authors' submission habit, the productivity of the institutes and the attractiveness of a journal toward potential authors.

Method and Tool

Co-occurrence analysis and social network analysis were used extensively in the current studies. Co-occurrence means that the institute's author published an article in the some journal. Social Network Analysis (SNA) can be considered vital for an understanding of this complex and dynamic structure. The major advantage of SNA method is that they can work at both micro and macro levels in their analysis of relational structure of different objects of study (David Mingguillo, 2010). The 2-mode data reflects the relationship between two nodes. In this study, the 2-mode network included two kinds of information: the institute and the journal. Social network analysis was performed by UCINET. 2-mode network is the network between the individual and organization in the social network analysis. It is based on the duality existed the individual and the organization or groups (LIU, 2009). 2-mode data can express by a rectangle matrix and it can convert into two 1-mode data. In this study the 2-mode data included the institute and the journal. The 2-mode data can express using the 2-mode network by the UCINET. The original matrix was normalized according the average papers' number.

Data Collection

The data were retrieved from the Chinese Social Sciences Citation Index (CSSCI). There are 53661 papers published in the 17 journal and there are 3898 first authors' affiliations during 1998 and 2011. The data were divided into three windows which were 1998-2003, 2004-2007, 2008-2011 according to the number of articles. The institutes were the top 25 institutes every window. At last the paper studied 17 journals and 38 institutes. The journals could be divided into 3 types: the library science journal, the information science journal

and the two fields mixed journal. The institutes included 3 public libraries, 2 research institutes and 33 universities.

Results and Discussion

The change of the institute-journal 2-mode network

The data findings show that there are more institutes published articles in the information science journals and the institutes are relatively less in the library science journals from 1998 to 2003 and from 2008 to 2011 from the Fig.1, Fig.2 and Fig.3. On the contrary there are more institutes published library science articles in the library science journals from 2004 to 2007.

Wuhan University, Nanjing University, Peking University and Nankai University are balanced published their articles in the library science and the information science journals from 1998 to 2003. Lately Sun Yat-Sen University and National Science Library of the Chinese Academy of Sciences entered the first group. Nanjing Agricultural University, Zhengzhou University, East China Normal University and Beijing Normal University published their works in the 4 mixed journals. It illustrates that these institutes' researchers are interesting both in the information science and library science. But their productive capability is lower than the first group.

The Institute of Scientific and Technical Information of China and similar institutions are the important information science institutes and the National Library of China and same kinds organizations are the important library science institutes.

Some institutes' submission shows regional characteristic obviously. For example, the Tianjin Library and Nankai University published more articles in the *Library Work and Study*. The Wuhan University published more articles in the *Document, Information & Knowledge*. The National Science Library of Chinese Academy of Sciences published more articles in the *Library and Information Service*. The Jilin University published more articles in the *Information Science*.

The *Library* and the *Library Development* are isolated in the Fig.3 it means the publications of the institutes' researcher are poorly published in the two journals. It reflected that their attractive ability is relatively weak.

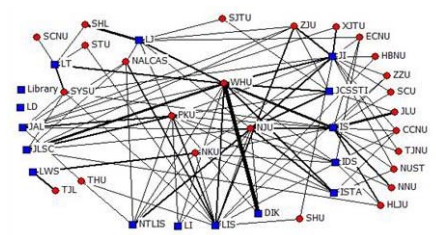


Fig.1 1998-2003 2-mode Network

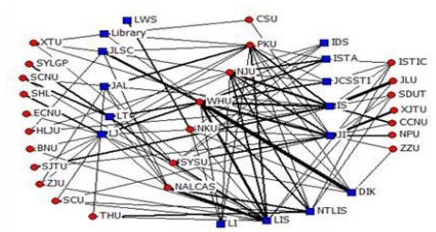


Fig.2 2004-2007 2-mode Network

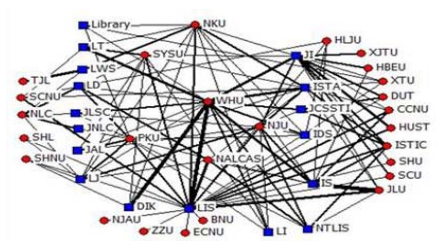


Fig.3 2008-2011 2-mode Network

The quantitative Analysis of the institutes' Centrality

It is more complex of the 2-mode networks' centrality. Degree, betweenness and closeness are 3 indexes to measure the centrality of the 2-mode network. From the data we can find there is more relativity between degree and closeness.

In this study the degree of the institute means its capability published articles in the some journal. The value of the degree, betweenness and closeness are changed in different window. The institutes can be divided into three parts in some window according to the 3 index. Wuhan University etc. are the first class, Shanghai University etc. are the second class and Xi'an Jiao tong University are the third class. The first class and the second class are balanced in the Information and Library science and the third class is asymmetrically outstanding in the Information science or in the Library science. The situation is changed by the time. For example, NALCAS is the second class from 1998 to 2003 and it is the first class from the 2004 to 2011.

The 17 journals can be divided into 3 parts: Library and Information Service, Information Science, Library Work and Study and so on. The value of the *Journal of Library Science in China* is relatively poor because their numbers of articles is less. We should synthetically treat the value according the realistic situation when we analysis the data.

Conclusion

Based on the institute-journal 2-mode network, we could find the relationship between the institutes and the journals form the quantity and quantity. We could divided the institutes and the journal into 3 parts according the 3 index. It also reflects the feature of the institutes and the journals.

Acknowledgments

This study supported by the Humanities and Social Science Funds of the Ministry of Education of the People's Republic of China (Project Code: 11YJC870024).

References

David Mingguillo(2010). Toward a new way of mapping scientific fields: Authors' competence for publishing in scholarly journals. Journal of the American Society for Information Science and Technology,61(4),771-786

Liu,Z.H.,Zheng,Y.N.(2011).On research specialty evolution mapping and its application. Journal of The China Society For Scientific and Technical Information,30(11),1178-1186

Ma,R.M.,Ni,C.Q.(2012). Author Coupling Analysis: An exploratory study on a new approach to discover intellectual structure of a discipline. Journal of Library Science in China,38(198),4-11

Qian,L.J.,Zhang,X.M.,Zheng,Y.N.(2008). Study on Appearance of Co-occurrence in Library and Information Science Institutions Overseas, Library and Information Service,52(11),49-52

Sun,H.S.(2012). Author keyword co-occurrence network analysis: an empirical research. Journal of Intelligence,31(9),63-67

Yang,L.B., Yang,L.Y.,Qiao,Z.H.(2010). A Study of Research Institutions' R & D Activities in Genomics Fields, Library and Information,1,93-98

Zhang,Z.L.Zhang,Z.Q.Li,X.Y.(2011).Co-occurrence analysis between research institutes and keywords based on 2-mode network, Journal of The China Society For Scientific and Technical Information, 30(12),1249-1260

COST ANALYSIS OF E –JOURNALS, BASED ON THE SCIENTIFIC COMMUNITIES USAGE OF SCIENCE DIRECT ONLINE DATABASE WITH SPECIAL REFERENCE TO BANARAS HINDU UNIVERSITY LIBRARY, INDIA

Dr.M.Anand Murugan; Dr.Ajay Kumar Srivastav

India,am9996@yahoo.com
Deputy Librarian,Banaras Hindu University,Varanasi,

Introduction

Library services and products have associated costs, including direct monetary costs and indirect costs such as time. The decision to acquire or provide a particular product or service should involve an examination of its costs and benefits to library customers. To assess the increasing prices of e-journals, we must find a way to compare journals with different amounts and quality of content, publishers, and subject-matter. Factors such as time, tangential costs such as paper or ink cartridges (or any other somewhat "hidden" costs), costs for training and materials, or any other factors that add to the cost of providing a service or product are considered indirect costs. Measuring benefits in a not-for-profit environment can be even more difficult. As part of the study the researcher decided to examine the three years science e direct online database downloads and subject wise downloads of articles. This article describes a study in which Cost Analysis was used to examine the cost-effectiveness of an electronic database.

Why Cost Analysis?

It is important to consider users and their demand for a particular e- journal. Who will use this e- journal? How often will they use it? E-journals further complicate the picture with complex pricing structures, online searching, hyperlinks, and server reliability. Regardless, the same problem of comparison remains: what a publisher charges for a particular journal does not necessarily reveal anything about that journal's relative value.

Science Direct Online Database

Elsevier Science Direct is the world's largest full-text database in the field of scientific, technical and medical (STM) information. Science Direct offers libraries and scientists globally have access to over eight million full-text articles - more than a quarter of the worlds electronic STM information - from over 1700 peer-reviewed journals, an expanding suite of Back files (titles loaded from Volume 1, Issue 1), as well as a growing range of authoritative books, including reference works, handbooks, book series and e-books.

In India, Science Direct is accessed at over 700 institutions - universities, corporate R&D centers, engineering and

medical colleges, governmental research laboratories etc

Usage Analysis by No. of Downloads

BHU Library approaches the Science Direct for getting the three years usage data. They received 3 years full text downloads details from Science Direct. The following figure arrived based on the science direct full text articles download. Three years includes 2010, 2011 and January to September 2012.

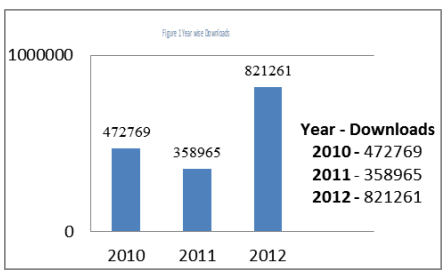


Figure 1. Year Vs. No. of Downloads

The above Figure 1 shows that the year wise downloads of full text science direct article. It clearly shows that Usage and downloads of articles in the year of 2010 to 2012. It is revealed from the figure that 4, 72,769 full text articles were downloaded in the year 2010, But, it was reduced in the year 2011, 3, 58,965 full text articles downloaded. Compare to the year 2010 with 2011, it was decreased. Its depends on many factors. But in the year 2012 Banaras Hindu University subscribed 13 subject collections through *2012 Elsevier & BHU Science Direct Agreement.* It increased the usage of the Science Direct articles. From January to September 2012 totally 8, 21, 261 full text articles are downloaded. It may increase in December 2012. This result shows that 23 subject collections which are subscribed through *2012 Elsevier & UGC-INFLIBNET* and *2012 Elsevier & BHU Agreement* is very useful to the users of BHU.

Usage Analysis by Subject Collection

BHU Library approaches the Science Direct for getting the subject wise usage data. Science Direct given the usage data during January to September 2012 was derived from 13 subject collections subscribed by BHU in 2012.

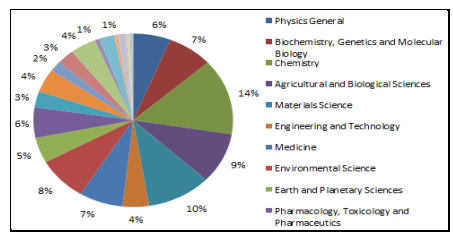


Figure 2. Usage Analysis by subject collection

This Figure 2 states that subject wise full text articles download. This data shows that the expected full-text downloads at Banaras Hindu University for the full-year of 2012 is 470,000 articles download which is an increase of 31% over the full-year usage for 2011. This figure 2 shows the usage and research output break-down by various subject collections. It is evident that a significant contribution (53%) to the usage on Science Direct during Jan-Sept 2012 was derived from 13 subject collections subscribed by BHU in 2012. This usage statistics is an indicator of high preference of the Elsevier journals (Science Direct) by the researchers of Banaras Hindu University.

Cost of per download

The researcher aims to identify the cost of per download of the article. Following formula retrieved the cost of per article. The no. of download divided by total subscription of the particular e-journal package gives the cost of per

download.*Cost of per download = Number of download / Total Cost of the subscription.*

No.	**Year**	*No. of download*	*Cost of per download Rs.*	*In $*
1.	*2010*	*472769*	*20.56*	
2.	*2011*	*358965*	*27.08*	
3.	*2012*	*821261*	*11.84*	

The above table shows in the year 2010, no. of downloads was 472769, cost of per article download is rupees 20.56, it was increased in the year 2011, due to decreasing the no. of downloads, so cost per article is 27.08. But it was dramatically decreased in the year 2012, due to overwhelming downloads, cost per article is 11.84 rupees. In this table shows that no. of downloads per year decides the cost of the article.

Conclusion

Results of the study suggest that Science Direct e-journals do affect use and cost of per article download to a significant degree. Most discrete analysis by subject collection will provide the locally useful information for collection development and research output of the subject. In order to accurately assess total costs of electronic journals, however our cost model should provide some visibility for annual "administrative", "access", or "platform" fees charged by a vendors like Science Direct. The formula for including these were approximated based on ratios of subscription costs among BHU for the titles in the study.

References

Colin K. Mick, "Cost Analysis of Information Systems and Services," *Annual Review of Information Science and Technology,* Vol.14 (1979) pp. 37–39

Tenopir, Carol and Donald W. King. "Towards Electronic Journals: Realities for Scientists, Librarians, and Publishers" Washington, DC: Special Libraries Association, 2000.

Montgomery, Carol Hansen."Measuring the Impact of an Electronic Journal Collection on Library Costs." D-Lib Magazine, Vol. 6 No. 10 Measuring costs of electronic journals Harter, Stephen P. and Charlotte E. Ford."Web-based

Analyses of E-journal Impact: Approaches, Problems, and Issues." *Journal of the American Society for Information Science*, Vol. 51 No. 13 (2000), pp.1159-1176.

A COVERAGE OVERLAP STUDY ON CITATION INDEX: COMMERCIAL DATABASES AND OPEN ACCESS SYSTEMS

Ming-yueh Tsay[1], You-min Lu[2] and Yen-chun Lin[3]

[1] *mytsay@nccu.edu.tw*
National Chengchi University, Graduate Institute of Library, Information Science and Archival Studies, NO.64,Sec.2,ZhiNan Rd.,Wenshan District, Taipei City 11605,Taiwan (R.O.C)

Introduction

The ways of knowledge dissemination have been changing with the evolution of media, especially in the last decade. Computers and webs have demonstrated great influences on the production, retrieval, dissemination and use of information, and new patterns of scholarly communication have been developed in the presence of citation systems, both commercial and open access. This study, therefore, aims to compare the above two kinds of citation systems to see their duplication, overlap, uniqueness and comprehensiveness.

Theories of coverage overlap among abstract, index and citation databases have been presented by researchers, and many evidence-based studies have been conducted ever since then. (Martyn,1967; Read & Smith, 2000). Gluck (1990) defined the journal coverage overlap as the ratio of the number of journal titles or articles in the intersection of two secondary sources to the number in their union. Comparisons between commercial databases and open access systems have drawn more attention as well after the widespread use of the Internet(Jacso; 2005; Bar-Ilan, 2010; and Rensleigh, 2011).

The studies on overlap and uniqueness can be very extensive, ranging from all sorts of data sources such as publishers, databases and search engines to all sorts of data types including journal articles, patents, web resources, and so on. The application of overlap studies may serve as references for libraries in selection of citation index databases.

Methods

Using the publications of the thirty A. M. Turing Award's Winners from 1990 to 2011 as the samples, the present study conducted retrievals in two commercial citation databases, such as SCIE, Scopus, and three open access citation systems of Google Scholar (GS), Microsoft Academic Search (MAS) and Citeseer^{X}. The bibliographic records retrieved were evaluated and cross-referenced to determine the comprehensiveness, overlap and uniqueness of the five citation databases according to the following steps.

1. Data Collection:

The names of the thirty Turing Award winners were used as queries for authors in each database to collect bibliographic records of their works from 1990 to 2012. To avoid confusions, background information about the thirty award winners was essential to identify the data collected; full texts were examined for further confirmation if necessary.

In case of any preclusion, no additional conditions were added to narrow down

the search results; records were manually filtered after exportation.

2. Data Refinement:

Co-written works by the thirty A. M. Turing Award winners were eliminated from the data collected. The refined data set was ready for analysis and comparison.

3. Data Analysis:

Cross-reference of databases by pair were conducted manually, and the five citation databases were paired as follows: SCIE-Scopus, SCIE-GS, SCIE-MAS, SCIE-CiteSeer^X, Scopus-GS, Scopus-MAS, Scopus-CiteSeer^X, GS-MAS, GS-CiteSeer^X and MAS-CiteSeer^X. Duplication within and overlap among databases were determined based on authors, year of presentation and content of texts.

Results

Types of samples retrieved from SCIE, Scopus, GS, MAS and CiteSeer^X include academic articles, conference proceedings, patents, unpublished manuscripts, course materials, memoirs, and institutional resources. Based on these samples, the results of duplicate within each database and comparison among the five databases were demonstrated and discussed as follows.

Intra-Duplication within Databases

Duplication may occur in each database for some reasons. Table 1 lists the number of records collected, number of records after refinement, number of duplicate, number of records without duplicate and percentage of intra-duplication. Take SCIE for instance, there are 1,188 records retrieved, which are reduced to 1,178 after data refinement, and the number of duplicates within the 1,178 records is 88, accounting for 8% of total records. Exclusive of the duplicates, 1,090 records remained is used in the examination of inter-database overlaps.

Table 1 Duplication within Each Citation Database

Database	(A)	(B)	(C)	(D)= (B)-(C)	(E)= (C)/(B)
SCIE	1,188	1,178	88	1,090	8%
Scopus	1,015	1,004	64	940	6%
GS	2,492	2,481	231	2,250	10%
MAS	2,488	2,479	236	2,243	10%
CiteSeer^X	902	897	49	888	5%

Note:
(A): Number of Records Collected;
(B): Number of Records after Refinement;
(C): Number of Duplicates;
(D): Number of Records without Duplicates;
(E): Intra-database Duplication

Table 2 Overlap among Citation Databases in Cross-reference Comparison

Databases by pair			(F)	(G)	(H)= (G)/(D)
1	SCIE-Scopus	SCIE	488	602	55%
		Scopus	388		64%
2	SCIE-GS	SCIE	548	542	50%
		GS	1708		24%
3	SCIE-MAS	SCIE	281	809	74%
		MAS	1434		36%
4	SCIE-CiteSeer^X	SCIE	811	279	26%
		CiteSeer^X	609		31%
5	Scopus-GS	Scopus	282	658	70%
		GS	1592		29%
6	Scopus-MAS	Scopus	177	763	81%
		MAS	1480		34%
7	Scopus-CiteSeer^X	Scopus	646	294	31%
		CiteSeer^X	594		33%
8	GS- MAS	GS	968	1,282	57%
		MAS	961		57%
9	GS-CiteSeer^X	GS	1800	450	20%
		CiteSeer^X	438		51%
10	MAS-CiteSeer^X.	MAS	1752	491	22%
		CiteSeer^X	397		55%

Note:
(F): Number of records exclusively indexed;
(G): Number of Identical Records;
(H): Overlap Percentage

Overlap among Databases: Cross-reference

Table 2 summarizes the number of records exclusively indexed in each database and the pair-wise overlap among the five databases under study.

For example, for the pair of SCIE-Scopus, 488 works by the thirty A. M. Turing Award winners are exclusively indexed in SCIE, accounting for 45% of total records; while 338 works are exclusively indexed in Scopus or 36% of total records. Both databases share 602 identical bibliographies, suggesting 55% overlap in SCIE and 64% in Scopus. In consideration of inter-database overlap, higher percentage implies lower quality, which indicates that SCIE surpasses Scopus in uniqueness. Open access citation systems excel commercial citation databases in general, while uniqueness of GS and MAS are better than CiteSeer^X.

Discussion and Conclusion

The results of this study reveals that the two general search engines, i.e., GS and MAS, are the most comprehensive and their number of records without duplication is more than twice of SCIE and Scopus. The comprehensiveness of CiteSeer^X is the poorest. In comparison of the two commercial citation databases, SCIE surpasses Scopus in comprehensiveness. The results also indicates that the intra-database duplication is the highest percentage of 10 for both GS and MS, while CiteSeer^X demonstrates the least intra-database duplication of 5%.

The study of overlap shows that the pair of GS-MAS demonstrates the highest number of identical number, while the pair of SCIE-CiteSeer^X the lowest. SCIE surpasses Scopus in uniqueness, but open access citation systems excel commercial citation databases in general, among which GS and MAS is better than CiteSeer^X.

Based on the findings of this study, citation data of each author, i.e. paper cited times, author cited times, author h index, etc. can be included and compared in further studies; whether or not the citation index services provide full-texts, and the number of accessible full-texts can be explored in future studies as well.

We hope the findings and suggestions based on the research results may serve as references for both commercial and open access service providers, and for libraries who consider to acquire or to build citation index systems on their own. Suggestions on indicators and tools for academic assessment are presented based on the comprehensiveness evaluation of each system as well.

Acknowledgements

This work was supported by grant NSC1007-2410-H-004-153-MY2 from the National Science Council, Taiwan, R.O.C.

References

Adriaanse, L & Rensleigh, C. (2011). Comparing Web of Science, Scopus and Google Scholar from an environmental sciences perspective. South African Journal of Libraries and Information Science, 77(2), 169-178.

Bar-Ilan, J. (2010). Citations to the "Introduction to Informetrics" indexed by WOS, Scopus and Google Scholar. Scientometrics, 82(3), 495-506.

Gluck, M. (1990). A review of journal coverage overlap with an extension to he definition of overlap. Journal of the American Society for Information Science, 41(1), 43-60.

Jacso, P. (2005). As we may search - Comparison of major features of the Web of Science, Scopus, and Google Scholar citation-based and citation-enhanced databases. Current Science, 89(9), 1537-1547.

Martyn, J. (1967). Tests on abstracts journals: Coverage overlap and indexing. Journal of Documentation, 23, 45-70.

Read, E. & Smith, C. (2000). Searching for library and information Science literature: a comparison of coverage in three databases. Library Computing, 19, 118-126.

FACTORS RELATED TO GENDER DIFFERENCES IN SCIENCE: A CO-WORD ANALYSIS

Tahereh Dehdarirad[1], Anna Villarroya[2] and Maite Barrios[3]

[1] *tdehdari@yahoo.com*
Department of Library and Information Science, University of Barcelona, Melcior de Palau, 140, 08014 Barcelona (Spain)

[2] *annavillarroya@ub.edu*
Department of Public Economy, Political Economy and Spanish Economy, University of Barcelona, Melcior de Palau, 140, 08014 Barcelona (Spain)

[3] *mbarrios@ub.edu*
Department of Methodology of Behavioral Sciences, University of Barcelona, Melcior de Palau, 140, 08014 Barcelona (Spain)

Introduction

The issues of gender mainstreaming, the role of gender in academic appointments and evaluation, and the participation of women in science as indicators of social and economic progress have attracted substantial attention from a broad array of researchers and national and international organisms (including, the Women in Industrial Research, 2003; the ENWISE (Enlarge Women in Science to East) expert group, 2004; She Figures, 2012; the WIRDEM (Women in Research Decision Making) expert group, 2006, and the EU-funded genSET project, 2010, among others). However, despite some progress, gender inequalities in science persist (EC, 2013).

A number of studies have sought to explain these discrepancies in various areas of science and academia by incorporating family-related factors, personal and institutional (structural) factors, professional factors, demographic and individual issues and factors related to disciplinary fields (Ceci and Williams, 2011; Fox, Fonseca, & Bao, 2011; Hunter & Leahy, 2010; Larivière, Vignola-Gagné, Villeneuve, Gélinas, & Gingras, 2011; Sax, Hagedorn, Arredondo, & Dicrisi III, 2002).

However, these studies fail to provide the kind of systematic and comprehensive overview of factors related to gender differences that might help guide future research and practices in the field. The aim of this study, therefore, is to undertake an analysis of the related literature using co-word analysis. Such an analysis helps visualize the division of a field into several subfields and the relationships that exist between them by providing insights into the evolution of the main topics discussed in the field over the years. Using co-word analysis, the present study aims to determine the structure of the knowledge network on the basis of the co-occurrence of terms, in order to describe the current state of the literature examining the factors that influence gender differences in science.

Method

The data set consists of a corpus of 651 articles and reviews, published between 1991 and 2012, dealing with factors related to gender differences in science. The data were extracted from the ISI Web of Science in February 2013, using a search that combined the principal terms related to the subject.

To carry out the co-word analysis, four sequential steps were followed: extraction and standardization of the keywords, construction of the co-occurrence matrix, clustering, and visual presentation of keyword groups. Author-provided keywords were extracted from papers. Keywords plus was used in those instances when no author-provided keywords were available.

Keywords and phrases were standardized manually and finally a total of 170 keywords were selected. In order to monitor the development of the scientific field, the results were divided into three periods, i.e. 1991-2001 (n=164 papers, 25.19%), 2002-2007 (n=147 papers, 22.58%), and 2008-2012 (n=340 papers, 52.23%).

The word-document occurrence matrix for each period was automatically built via SPSS v. 20. The resulting matrices were then exported to Ucinet v.6. In Ucinet the word-document matrix was transformed into a word co-occurrence matrix; the similarities between items were also calculated using the Jaccard similarity index. Hierarchical clustering analysis was then conducted via SPSS v.20 using Ward's method and the squared Euclidean distance was applied as the distance measure.

Based on the dendrogram generated by the clustering algorithm, clusters of keywords were derived. The clusters were then transformed into networks in Ucinet v.6. Finally, after calculating the density and centrality of each cluster, the keyword networks were displayed in a strategic diagram using Excel. It should be borne in mind that, in each strategic diagram, the volume of the spheres is proportional to the number of documents corresponding to each cluster.

Results

Based on the hierarchical clustering analysis, four clusters of keywords were identified for the first period (1991-2001), ten clusters for the second period (2002-2007), and finally, sixteen for the third period (2008-2012). A strategic diagram depicting the relative positions of each cluster was produced for each period to facilitate interpretation (Figures 1, 2 and 3).

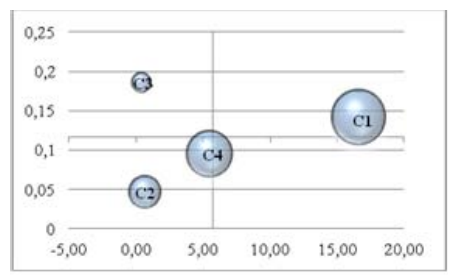


Figure 1. Strategic diagram. Period 1991-2001

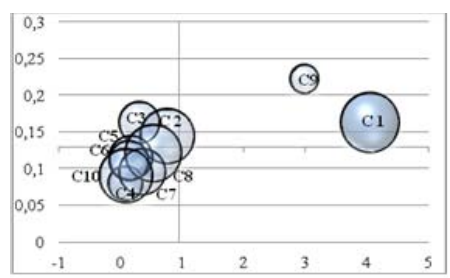


Figure 2. Strategic diagram. Period 2002-2007

Five motor-themes appeared in the upper-right quadrant of the diagrams on the basis of their high density and centrality. The motor-themes were:

"Gender inequalities in labor markets and universities" (cluster 1) in the first period, *"Career satisfaction in medicine"* (cluster 1) and *"Academic career in sociology"* (cluster 9) in the second period and finally *"Progression in academic medicine"* (cluster 9) and *"Staff composition and climate in academia"* in the third period (cluster 11).

Only two themes in the diagrams were present in all three periods: *"Mobility of women academics"*, and *"Institutional discrimination"*. Some themes emerged and were maintained in the periods that followed: *"Work-life balance in academia", "Racial inequality in higher education"*, and *"Progression in academic medicine"* appeared in both second and third periods. *"Promotion differences"* appeared in both first and second periods.

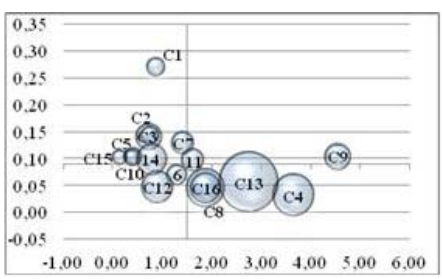


Figure 3. Strategic diagram. Period 2008-2012

Conclusion

The present results provide interesting insights into the evolution of the literature examining factors related to gender differences in science. An overall analysis of the three periods clearly revealed an increase in the number of themes in the most recent period (2008-2012) and a variety of motor-themes depending on the period studied.

References

Ceci, S. J.; Williams, W. M. (2011). Understanding current causes of women's underrepresentation in science. PNAS 108, 8, 3157-3162. http://www.pnas.org/cgi/doi/10.1073/pnas.1014871108

Enwise. (2003). EUR 20955 — Waste of talents: turning private struggles into a public issue. Women and Science in the Enwise countries. Luxembourg: Office for Official Publications of the European Communities.

European Commission. Directorate-General for Research and Innovation (2013). *She Figures. Gender in Research and Innovation*. Brussels: Directorate-General for Research and Innovation.

Fox, M. F., Fonseca, C., Bao, J. (2011). Work and family conflict in academic science: Patterns and predictors among women and men in research universities. *Social Studies of Science, 41*(5), 715–735.

Hunter, L. A., & Leahey, E. (2010). Parenting and research productivity: New evidence and methods. *Social Studies of Science*, 40(3), 433–451.

Larivière, V., Vignola-Gagné, E., Villeneuve, C., Gélinas, P., & Gingras, Y. (2011). Sex differences in research funding, productivity and impact: an analysis of Québec university professors. *Scientometrics*, 87(3), 483-498.

Sax, L. J., Hagedorn, L. S., Arredondo, M., & Dicrisi III, F. A. (2002). Faculty Research Productivity: Exploring the Role of Gender and Family-related Factors. *Research in Higher Education*, 43(4), 423-46.

WIR. (2003). Women in Industrial Research: a Wake Up Call for European Industry. Brussels: European Commission.

THE CROSSCHECK PLAGIARISM SYSTEM: A BRIEF STUDY FOR SIMILARITY

Edilson Damasio[1]

[1] *edamasio@uem.br*
Federal University of Rio de Janeiro-UFRJ-IBICT. Universidade Estadual de Maringá. Eduem. Av. Colombo, 5790 bloco 40, CEP 87020-900, Maringá, Paraná (Brazil). Rio de Janeiro-IBICT.

Introduction

The plagiarism identification in articles submitted for publication in scientific journals has been configured as an essential requirement for the decision of the scientific editors.

Among the different systems of identification of plagiarism, the CrossCheck has been used by leading scientific publishers in the world, because it allows them to identify similar documents on the Internet or in the others journals (fully or partially) those indexed in its a database. Editors of scientific journals countries, unaware or no use CrossCheck, is already used by reputable scientific journals and publishers.

From this premise, this paper aims to show the results of a survey *2012 CrossCheck Survey,* used by CrossRef to feedback to the scientific journals editors and publishers, about the use the SIMILARITY in a new submissions.

The objective of the article is compared the results from the journals publish in the scientific community, to the percentages of plagiarism identification in the CrossCheck Survey.

Was identified four journals with published the results of similarity reports. The Web Of Science and Scopus it's database to results of journals with articles with CrossCheck percentage of plagiarism articles.

The low index of journals that publish your similarity results, it's related to the use of the tools by the editors. The similarity it's a support do decision making by identifying plagiarism and scientific misconduct.

The four journals, already using CrossCheck, to identify the use of the system is essential for the serial publication with identification of plagiarism policies defined.

The title of the journals will not be identified, and verification period is the year 2004 to 2012. The results will be distributed in journals A B C D.

CrossCheck

The CrossCheck is a system developed in cooperative form to the members, publishers, journals and your contents. All affiliates must enable your will be part of the database.

Aimed mainly to the demand of plagiarism and misconduct identified by editors of scientific journals. The major publishers of the world participate in the database in order to have a most possible content to area of knowledge.

The system is also offered to members of CrossRef, which should identify the use of the DOI (Digital Object Identifier) redirection to textual content published.

Plagiarism and Scientific Journals

The plagiarism utilization and ethical aspects of scientific communication

have always known by the editor, researchers, teachers and students, and difficult to identify. All these actors are possible and require users to identify plagiarism system, to assist them in comparing similarity (Fig. 1).

The Elsevier director of services Catriona Fennell cited by Butler (2010) say with, "Establishing plagiarism requires 'expert interpretation' of both articles says Fennell. The software gives an estimate of the percentage similarity between a submitted article and ones that have already been published, and highlights text they have in common. But similar articles are sometimes false positives, and some incidents of plagiarism are more serious than others".

Similarity

The scientific literature it's universal and constant increased. Identify the content for the publication of an article and accepted by publishers, researchers, all actors and us scientific production. Scientific journals that can't error and publish similar texts, results, and indirectly take plagiarism of texts, showing errata and retractions.

An author who wants to identify if others are quoting their texts, this mapping is very important to work on their lines of research and compare their scientific production. Identification similarity systems are essential in academic area.

The similarity is studied in scientometrics and informetrics, the words as objects in databases and recovery variables and measure to the relevance and percentages (number of characters or words) similar or with minor changes (Bornmann et al., 2008).

Figure 1. Scientist and words shake. Font: Nature comment (2012).

Results

In table 1 the results of brief research with four journals in Web Of Science and Scopus. This percentage it's published by the editors. The scientific production about similarity and CrossCheck users, it's very small.

Table 1. Journals and percentage of similarity with CrossCheck. Font: Web Of Science / Scopus

Journal	Total n	N	%
A	216	21	10
B	56	13	23
C	na	na	10
D	na	na	31

CrossCheck Survey Results 2012

It's presented a brief and principal dates to the objective the poster. This study it's available and identifies various possibilities of the results comparations.

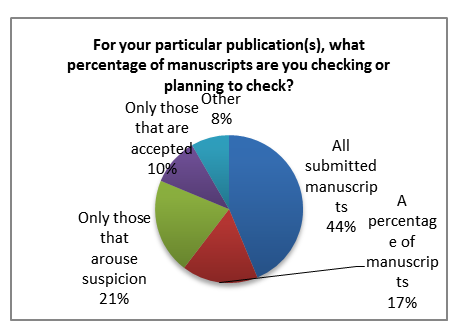


Figure 2. Percentages. Font: CrossCheck Survey (2012)

Compared with the group of four journals the average result of similarity was 32%, so this result is not definitive percentage, due mainly to a percentage of publishers not characterize that is 10 or 99% similarity was a plagiarism. Thus the similarity measure is one to be look at by the publisher and your objectives, focus, area, but the decision depends on textual content which will be similar to that seen by the human eye.

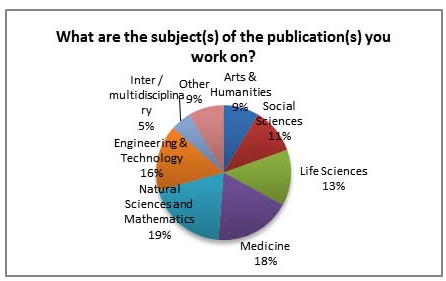


Figure 3. Subject areas. Font: CrossCheck Survey (2012).

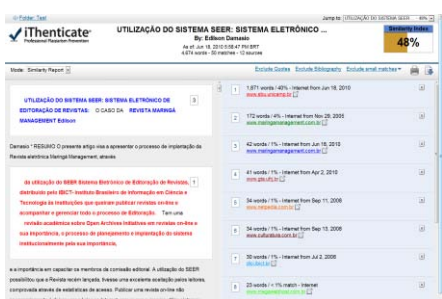


Figure 4. CrossCheck similarity report results with percentage and link to the results. Font: Damasio (2012).

Conclusion

The CrossCheck it's utilized by Elsevier and Nature and others best publishers, is a large partners to scientific databases. In actual numbers is utilized by 300 publishers, with 70.000 journals titles, 40 million of indexed contents, and 30.000 documents checked monthly. There are methods that content posted on this system has security and legal guarantees. The service is offered a low cost, a small annuity and checking documents at a cost of $0,75 US dollars. The system is a useful tool in numerous processes in institutions, such as checking others documents, theses, dissertations, projects, monographs and books have other costs. The content of database is that articles and collections to the all journals, and the journal editors should use it mainly for verification of newly submitted articles to journals.

Acknowledgments

Acknowledgments to State University of Maringa, to Federal University of Rio de Janeiro to possibility of research in this area (Information Sciences) and to CrossRef for the data numbers.

References

Bornmann, L., Mutz, R., Neuhaus, C.& Hans-Dieter, D. (2008). Citation counts for research evaluation: standards of good practice for analyzing bibliometric data and presenting and interpreting results. *Ethics in Science and Environmental Politics*, 8, 93-102. doi: 10.3354/esep00084

Butler, D. (2010). Journals step up plagiarism policing. *Nature*, 466, 167.

CrossCheck (2012). November 2012 CrossCheck Survey Results. Oxford: CrossRef.

Damasio, E. (2012). CrossCheck: identificação de plágio por similaridade. *Encontro Brasileiro de Bibliometria e Cientometria*, 3.

Nature comment (2012). How to stop plagiarism. *Nature*, 481, 7379, 21-23.

Zhang, Y. (2010). Chinese journal finds 31% of submissions plagiarized. *Nature*, 467. doi: 10.1038/467153d

CUMULATIVE CAPABILITIES IN COLOMBIAN UNIVERSITIES: AN EVALUATION USING SCIENTIFIC PRODUCTIVITY

Sandra Carolina Rivera-Torres[1], Marcela Galvis-Restrepo[2] and Jenny Cárdenas-Osorio[3]

[1]*crivera@ocyt.org.co;*
Human Resources Indicators of Sciences and Technology, Observatorio Colombiano de Ciencia y Tecnología, Carrera 15 Nº 37-59, Bogotá, Colombia; Universidad Nacional de Colombia

[2] *mgalvis@ocyt.org.co;* [3]*jcardenas@ocyt.org.co*
Human Resources Indicators of Sciences and Technology, Observatorio Colombiano de Ciencia y Tecnología, Carrera 15 Nº 37-59, Bogotá, Colombia

Introduction

Research was institutionalized in Colombian universities around fifty years ago with the creation of government institutions to stimulate this activity, and the support of loans from the Inter-American Development Bank –IDB- that were used to start building scientific and technological capabilities (Bucheli et al., 2012; CINDA, 2012; OECD & The World Bank, 2013). The institutional development of the science, technology and innovation system, gives priority to research groups as organizational units for research; a structure largely promoted by Colombian Administrative Department of Science, Technology and Innovation –Colciencias-, the government body in charge of measuring research group's scientific capabilities and assigning resources according to their performance (Villaveces & Forero, 2007; OCyT, 2012).

Research groups as organizational units were adopted by universities and research institutions as well. Some universities have even developed their own financing schemes intended to stimulate their growth and sustainability (Colciencias, 2008; OCyT, 2012). In 2011-2012, the Colombian Observatory of Science and Technology –OCyT- developed a methodology to evaluate the research groups of one of Colombia's leading universities, which can be used for tracking and impact evaluation of R&D policy.

As a result, in this work we focus on scientific capabilities of research groups measured by their researchers' scientific production, as defined by Colciencias' model. In the previous evaluation (OCyT, 2012), we found evidence that the production is highly concentrated in a few researchers following Zipf's power law in an eight year time frame (Sutter & Kochner, 2001). This approach allows for a representation researcher's capabilities, taking into account the accumulated knowledge in the publication trajectory of the population.

Methodology

We looked at publications registered in ScienTi[179] database associated to researchers working in research groups in the incumbent university; afterwards, we applied concurrency algorithms to match those papers, with publications in ISI- Thomson Reuters Web of Science database.

Additionally, since the knowledge codified in publications is an evolutionary variable that can be accumulates over time; in this analysis we use cumulative distributions in order to determine the future trend of scientific publications as means to understand the individual's capabilities (Bozeman & Corley, 2004; Lepori & Barré, 2008).

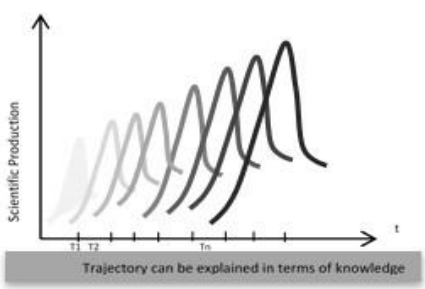


Figure 1. Capabilities Measurement Using Scientific Production

According to Abe & Rajagopal (2001) the state of the scientific production of a group of researchers fits the Cauchy-Lorentz distribution. This means that the difference in productivity p, between the members (n) of the research group can be modelled with a Lorentz distribution, given by the equation:

$$p(n) = \frac{1}{\pi}\frac{\gamma}{(n-n_c)^2+\gamma^2}, \qquad (1)$$

where n_c is the arithmetic half of the population and γ is the half-width. As shown in Figure 1, the distribution is useful to evaluate the behaviour of the production of researchers each year and by accumulating it; we can observe the degree of concentration of the knowledge generated.

First, we accumulate the production of each researcher in a given scientific field in 2002-2008. In Figure 2 we show an example for a population of 167 researchers located in the horizontal axis in the field of medical sciences of the Universidad de Antioquia (OCyT, 2012); production is represented in the vertical axis.

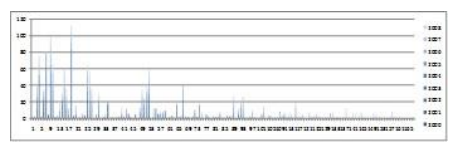


Figure 2. Accumulated Scientific Production in 2000-2008, Researchers in Medical and Health Sciences

Then we organize the results locating the most productive researcher in the centre; the second and third most productive are located to the right and left and so on. Figure 3 shows the data in Figure 2 organized in a centralized way. These data are shown as an example for a small population of researchers. In the case of Colombian universities, this tool is useful to describe the trajectory of the research groups.

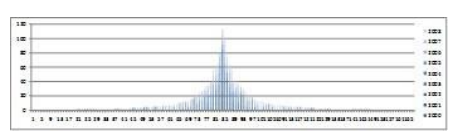


Figure 3. Centralized Accumulated Scientific Production in 2000-2008, Researchers in Medical and Health Sciences

[179] The Colombian platform where research groups and researchers register their information in order to get recognition and resources from Colciencias; it contains information on researcher's curriculum vitae (CvLAC) and research groups (GrupLAC)

As shown in Figure 4, production is highly concentrated on a few researchers who have the highest share of the scientific production. In this graph we fitted a Lorentz distribution to the centralized data.

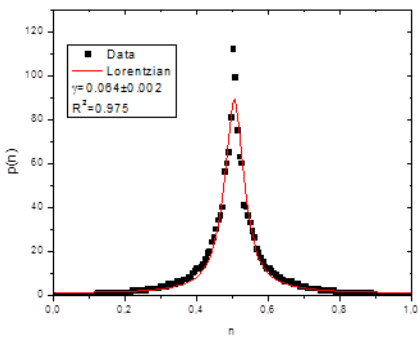


Figure 4. Lorentz Fitting of the Centralized Distribution

Using Lorentz distributions, the effect of a researcher in the population can be defined as its percentage contribution given by the relationship:

$$E(n) = \frac{p(n)}{P_T}, \quad (2)$$

Usually the centre value is near N/2, so it would be enough to find the number γ (the half width) to determine the distribution of the production in the population. A property of the Lorentz distribution is:

$$P_T = 2\int_{n_c-\gamma}^{n_c+\gamma} p(n)dn, \quad (3)$$

So the range [n_c-γ,n_c+γ] (with size 2γ) contains half of the population. With this result we can define the percentage of the population concentrating 50% of the production as:

$$N_m = \frac{2\gamma}{N}, \quad (4)$$

This result can be used as a measure for the concentration of production in the population of researchers, and eventually for research groups' comparisons.

Conclusions

The methodology presented here is useful for evaluating variables reflecting the knowledge accumulation observed in R&D activities. An additional application consists in mapping the trajectory of different groups like individuals, research groups or institutions; additionally it is useful to observe the degree of development reached by each individual in the S&T system and use it to target specific populations of highly achieving individuals. Specifically in the Colombian context, the methodology helps to measure capacity while it is understood that individuals have an evolutionary trajectory, and they shape a growing and heterogeneous community.

References

Abe, S., & Rajagopal, A. (2001). Information theoretic approach to statistical properties of multivariate Cauchy-Lorentz distributions. *Journal of Physics A: Mathematical and General, 34*(42), 8727.

Bozeman, B., & Corley, E. (2004). Scientists' collaboration strategies: implications for scientific and technical human capital. *Research Policy, 33*(4), 599-616.

Bucheli, V., Díaz, A., Calderón, J. P., Lemoine, P., Valdivia, J. A., Villaveces, J. L., et al. (2012). Growth of scientific production in Colombian universities: an intellectual capital-based approach. *Scientometrics, 91*(2), 369-382.

CINDA. (2010). Educación superior en Iberoamérica Informe 2010.

Lepori, B., Barré, R., & Filliatreau, G. (2008). New perspectives and challenges for the design and production of S\&T indicators. *Research Evaluation, 17*(1), 33-44.

OCyT. (2012). *Indicadores de Ciencia y Tecnología Colombia 2012.*

OECD, & Bank, T. W. (2012). *Reviews of National Policies for Education: Tertiary Education in Colombia 2012*. OECD Publishing.

Sutter, M., & Kochner, M. G. (2001). Power laws of research output. Evidence for journals of economics. Scientometrics, 51(2), 405-414.

A DESCRIPTIVE STUDY OF INACCURACY IN ARTICLE TITLES ON BIBLIOMETRICS PUBLISHED IN BIOMEDICAL JOURNALS

Rafael Aleixandre-Benavent[1], Vicent Montalt[2], Juan Carlos Valderrama-Zurian[3], Miguel Castellano-Gómez[4]

[1] *aleixand@uv.es*
IHMC López Piñero, CSIC-Plaza Cisneros 4, 46003-Valencia (Spain)

[2] *montalt@trad.uji.es*
Departament de Traducció i Comunicació. Universitat Jaume I.Campus de Riu Sec. 12071 Castelló (Spain)

[3] *juan.valderrama@uv.es.*
Instituto de Documentación y Tecnologías de la Información. Universidad Católica de Valencia. C/ Quevedo, 2. 46001-Valencia (Spain)

[4] *castellano_mig@gva.es*
Hospital Aranu de Vilanova. Conselleria de Sanitat. Genralitat Valenciana (Spain)

Introduction

Although it is a very small part of the research paper, the title plays an important role and certain pragmatic functions as the first point of contact between writer and potential reader: to provide a general and brief description of the content of the article, to attract attention evoking interest in the reader, to inform, and sometime, to startle (Haggan, 2004). Readers generally decide whether to read an article or not by seeing the title first. For this reason, titles in science mirror a set of requisites that are crucial to the construction, communication, and progress of new knowledge (Cheng, Kuo and Kuo, 2012; Soler, 2007). Many titles do not comply with the standards established in style manual on scientific writing by using sensationalist devices such as interrogation marks, exclamation marks, metaphors, double meanings and vague expressions in order to catch the readers' attention. The purpose of this article is to analyse the lack of accuracy of titles in articles on bibliometrics published in biomedical journals.

Methods

We analysed a corpus of 1.100 titles included in PubMed database between 2009 and 2011 and retrieved under the major MeSH topic "bibliometrics". Different types of inaccuracy were identified and classified using an explicit typology developed for this particular study.

Results

24,7% of the titles contain some type of inaccuracy. Editorial titles show a higher percentage of inaccuracy occurrences (11.71%) than original articles (9.28%) and letters (3.7%). The most frequent type of inaccuracy is including a question in the title, which amounts for 32.12% of the papers (table

1). The next category down is vague and imprecise expressions (18.98%) (table 2), acronyms (14.96%) and double meanings (13.88%) (table 3). Some titles also use amusing titles, biblical sentences and movie titles (table 4).

Table 1. Examples of topic-question titles

H-index-a good measure of research activity? (Tidsskr Nor Laegefor.2011;131:2494-6).
What are we reading now? An update on the papers published in the orthodontic literature (1999-2008). J Orthod. 2011;38:196-207.
How to judge a book by its cover? How useful are bibliometric indices for the evaluation of "scientific quality" or "scientific productivity"? Ann Anat. 2011;193:191-6.
Ranking hepatologists: which Hirsch's h-index to prevent the "e-crise de foi-e"? Clin Res Hepatol Gastroenterol. 2011; 35: 375-86.
What does the Journal's impact factor mean to you? J Am Diet Assoc. 2011;111:41-4.

Table 2. Examples of vague expressions

Impact and scholarship. J Nurs Scholarsh. 2010 Sep 1;42(3):233.
Journal Impact Factor: it will go away soon. Clin Chem Lab Med. 2009;47(11):1317-8.
Spreading the word. Nurs Stand. 2009;23:22-3.
Ideas with impact. Nurs Inq. 2011;18:277.
A time of change. J Hum Nutr Diet. 2011;24:1-2.

Discussion

Writing the titles to scientific articles is a challenging exercise that demands the use of various skills. Academic writing textbooks and style manuals have proposed the elements of good research articles titles (Cheng, Kuo and Kuo, 2012; Swales and Feak, 2004): (1) The title should indicate the topic of the study. (2) The title should indicate the scope of the study. (3) The title should be self-explanatory to readers in the chosen area". Day and Gastel (2006) defined a good title as "the fewest possible words that adequately describe the contents of the paper". The journals do not often provide rules for writing the titles in its guidelines for manuscript submission. Moreover, the guidelines known as the Uniform Requirements for

Table 3. Examples of double meanings expressions

Watching the river flow. (Rev Port Pneumol. 2011;17:197-8).
Impact factor: vitamin or poison? (Sao Paulo Med J. 2010;128:185-6).
Staying on the cutting edge. (Augment Altern Comm. 2010; 26:223-5).
The race for the impact factor. (J Sleep Res. 2009;18:283-4).
The road is made by walking. (Gac Sanit. 2010;24:1-4).
The impact of the impact factor. (Can J Urol. 2009;16:4445-6).
The first and the last will become the best (Orv Hetil. 2010;151:1236-7).
Impact factor: does it have an impact? (J Ayub Med Coll Abbot. 2009; 21: 180).

Table 4. Examples of biblical sentences and movie titles

But many that are the first shall be last, and the last shall be first. (FASEB J. 2009;23:1283).
Impact factor wars: Episode V-the empire strikes back. (J Child Neurol. 2009;24:260-2).
Looking back to the future. (Worldviews Evid Based Nurs. 2009; 6:1-2).
The impact factor for evaluating scientists: the good, the bad and the ugly. (Clin Chem Lab Med. 2009;47:1585-6).

Manuscripts submitted to Biomedical Journals, or Vancouver style, also provide limited information about how the tiles of papers are to be written.

Topic-question titles can stimulate reader's interests by the use of a question. The question title construction seems to allow authors the possibility of posing questions on such object as an indication that, in spite of the current state-of-the-art about it, there are, still, queries in need of reply, interpretation, and conclusion (Soler, 2007).

Since the use of a metaphor can greatly arouse reader's curiosity, the juxtaposition of a metaphor or a double meanings sentence with the real research topic in a compound title seems a clever construction that can attract readers to think about the association between them. For instance, when readers read "The race for the impact factor", they may be puzzled, but attracted by the metaphorical expression of "the race", as they read the other part of the title, "the impact factor", which reveals the research topic, they realize what is implied in the metaphor. The use of this sort of construction can often make a strong impression on readers.

Similar to metaphor-topic titles, the use of humor in scientific titles makes sense if we take the point of view of the title as a persuasion tool for attracting readers, but in general, decrease the tendency to read an article and treat its contents seriously (Hartley, 2007).

Conclusions

In order to make scientific articles more effective, most titles of scientific articles on bibliometric topics use a variety of devices, even if they do not comply with the conventions of scientific writing. We recommend that authors write more descriptive titles that accurately reflect the content of the articles so that readers can understand them better and retrieve them better from bibliographic databases. It would be useful in future research to ask authors about their practices in choosing titles when writing papers singly or with others.

References

Cheng SW, Kuo CW, Kuo CH. Research article titles in applied linguistics. Journal of Academic Language & Learning Vol. 6, No. 1, 2012, A1-A14.

Day, R. A., Gastel B (2006). How to write and publish a scientific paper. Westport CT: Greenwood Press.

Haggan, M. (2004). Research paper titles in literature, linguistics and science: Dimensions of attraction. Journal of Pragmatics, 36(2), 293-317.

Hartley J. There's more to the title than meets the eye: Exploring the possibilities. Journal of Technical Writing & Communication. 2007; 37:95-101.

Soler, V. (2007). Writing titles in science: An exploratory study. English for Specific Purposes, 26(1), 90-102.

Swales, J. M., & Feak, C. B. (1994). Academic writing for graduate students: Essential tasks and skills (1st ed.). Ann Arbor, MI: Univ. Michigan Press.

DIFFUSION OF BRAZILIAN STATISTIC INFORMATION

Sônia Regina Zanotto[1], Samile Andréa de Souza Vanz[2] and Ida Regina Chittó Stumpf[3]

[1] *zanotto.sonia@gmail.com*
IBGE-RS, Av. Augusto de Carvalho, 1205, Cep 90010-390 – Porto Alegre, RS (Brazil)

[2] *samilevanz@terra.com.br*
Federal University of Rio Grande do Sul, Information Science Department, Post-Graduation Program on Communication and Information, Rua Ramiro Barcelos 2705 sala 214, Cep 90035-007 – Porto Alegre, RS (Brazil)

[3] *irstumpf@ufrgs.br*
Federal University of Rio Grande do Sul, Information Science Department, Post-Graduation Program on Communication and Information, Rua Ramiro Barcelos 2705 sala 214, Cep 90035-007 – Porto Alegre, RS (Brazil)

Introduction

This poster brings analysis on citations received by IBGE – Brazilian Institute of Geography and Statistics – publications from the perspective of the diffusion factor calculation theory and methodology. Considering that citations represent a way to measure how scientific ideas spread, we have sought ground in the diffusion factor theory proposed by Rousseau, Liu and Ye (2012) to present a practical application.

IBGE has been the official Brazilian agency in charge of public statistic information since 1938; among other documents it publishes the Census and the National Household Sample Survey (PNAD). The citations received by documents produced by the Institute are spread among dozens of scientific journals, published in different continents and in several languages. This fact in itself indicates that the publications are properly disseminated (Zanotto, 2011; Zanotto, Vanz & Stumpf, 2011).

Bibliometric indicators are necessary in bringing new perspective for the understanding of scientific communication (Frandsen, Rousseau & Rowlands, 2006). There is consensus that they must be cautiously analyzed and interpreted within a context, since they are incomplete as measurement and they generally present isolated results. The impact measurement of citations that has been applied does not take into consideration who or what citing sources are or how the citations are dispersed, yet in a certain way the geographical reach of citing sources represents the extension of the geographical impact of the information. (Rowlands, 2002). We understand that it is necessary to show not only the impact, but also the dimension in which information is received by the community.

Journal Diffusion Factors have been introduced in order to measure the influence on scientific research and the diffusion of journals, in an attempt to complement the Impact Factor (Rowlands, 2002). Since its

introduction, several researchers have been developing a measurement called diffusion factor by means of different data collection techniques (Frandsen, 2004). In Brazil, Rummler (2006) presented the Index of Segmental Dispersion, an indicator that can be applied to a tittle, an author, a journal or a field of knowledge, and considered the possibility of measuring the dimension of the extension of the impact of a single analysis unit, as it is applied in the citation analysis. More recently, Rousseau, Liu & Ye (2012) presented the concept that scientific ideas flow through a *layered system* and this is how diffusion must be measured. Based on the idea that citations represent the diffusion of knowledge and that the standard Gini coefficient reflects the measure of such diffusion, the Rousseau, Liu & Ye (2012) methodology was applied to citations on IBGE publications in the period 2001-2010 with the purpose of calculating its diffusion factor.

Methodology

IBGE scientific output citations were identified in the Web of Science. In the *Cited Author* field we used a search key made up of the group of different variant forms of the abbreviation and the complete name of the Institution in English and Portuguese for the period 2001-2010, without any restrictions on type of document to be retrieved. Cleaning and standardizing procedures for the authors' names and respective affiliated institutions were needed. Data were analyzed with the BibExcel software and Microsoft Excel 2007 program for electronic spreadsheets. We isolated information on authorship from the AU field and they were accounted for year by year, along with respective information on institution and country contained in the C1 field of address from the group of data retrieved. The count and fractioning followed the Rousseau, Liu & Ye (2012) methodology.

Results

We identified 3,158 documents citing IBGE scientific output in the period between 2001-2010. When analyzing the authors of these documents, we found 10,707 names, leading to a mean of 1.29 citations per author and frequency that ranged from 1 to 19 citations of IBGE publications per author in the period. The citing authors were affiliated to 1,272 different institutions. Among the 7,587 occurrences of countries in the institutional link of the authors, 6,168 (81.3%) of them referred to Brazil, 9% to the United States of America, 2% to England and the other occurrences were spread among 49 different countries. The distribution per continent is the following: South America (82.35%); North America (9.71%); Europe (6.93%); and Central America, Asia, Oceania, Africa and Middle East (approximately 1% of citing authors).

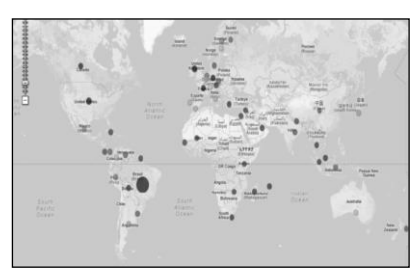


Figure 1 – Geographical distribution of countries to which authors who cited IBGE scientific output were affiliated to in the period between 2001-2010

Out of the 1,272 institutions to which citing authors were affiliated, 748 (58.80%) were Brazilian or based in Brazil, while 518 (40.72%) were foreign. Institutions with an educational focus lead (47.96%), followed by P&D

institutions (22.88%) and the others were from the public sector, such as State Secretaries, hospitals, agricultural business, livestock and service, besides industry, among others.

Table 1 – Gini index applied to citing authors, institutions and countries 2001-2010

Year	Number of Citing Articles (CA)	Fraction Sum Citing Authors (AU)	Fraction Sum Citing Institutions (UNI)	Fraction Sum Citing Countries (CO)	G_e
2001	106	105.49	70.77	8.50	0.89
2002	110	109.24	70.93	8.42	0.89
2003	**123**	**121.76**	**76.11**	**13.57**	**0.90**
2004	177	176.00	97.33	10.83	0.89
2005	184	182.25	91.27	10.33	0.89
2006	222	221.22	122.51	18.15	0.89
2007	447	444.96	185.47	19.08	0.89
2008	654	647.92	243.75	21.42	0.89
2009	**673**	**671.42**	**264.99**	**19.90**	**0.88**
2010	462	460.72	254.26	22.33	0.89

Source: Research data

With the purpose of proving that the citations reflect satisfactory diffusion of the knowledge produced by IBGE we applied the diffusion factors analysis proposed by Rousseau, Liu & Ye (2012). After applying G_e formula, we obtained quite satisfactory results, since the Gini index remained between 0.88 in 2009 and 0.90 in 2003, indicating a high dispersion factor, that is to say, there is a great number of authors, institutions and different countries that cited IBGE publications in the period, as can be seen in table 1. Although the absolute number of citations varies increasingly in the period, the G_e index remains in the same level in the decade under study.

Conclusions

The greater the number of authors, institutions and different countries, the higher is the degree of diffusion and this ratio is clear when analyzing the absolute citations to IBGE. By applying the diffusion factor, we were able to prove what the absolute numbers seemed to suggest: Gini index between 9.88 in 2009 and 0.90 in 2003 indicate a high dispersion factor, thus, there is a great number of authors, institutions and different countries that cite IBGE publications.

Much like Rousseau, Liu & Ye (2012) considered the study that defined the standardized Gini coefficient (Ge) preliminary; this study has shared experimental features and is subject to criticism and comments.

References

Rousseau, R., Liu, Y.& Ye, F.Y. (2012). A preliminary investigation on diffusion through a layered system. *Journal of Informetrics*, 6, 177-191.

Frandsen, T.F. (2004) Journal diffusion factors – a measure of diffusion? *Aslib Proceedings*, 56, 5-11.

Frandsen, T.F., Rousseau, R.& Rowlands, I. (2006). Diffusion factors. *Journal of Documentation*, 62, 58-7.

Rowlands, I. (2002). Journal diffusion factors: a new approach to measuring research influence. *Aslib Proceedings*, 54, 77-84.

Rummler, G. (2006). Modelagem de um indicador bibliométrico para análise da dispersão de conhecimentos. *Ciência da Informação*, 35, 63-71.

Zanotto, S.R., Vanz, S.A.S. & Stumpf, I.R.C. (2011). A informação estatística oficial produzida pelo IBGE e a sua difusão geográfica. In: *XII Encontro Nacional de Ciência da Informação e Biblioteconomia*, Brasília. (XII ENANCIB).

Zanotto, S.R. (2011). Informação estatística oficial produzida pelo IBGE: apropriação pela comunidade científica brasileira no período de 2001 a 2009. Dissertação de Mestrado, UFRGS, Porto Alegre, Brasil.

DISCOVERING AUTHOR IMPACT: A NOVEL INDICATOR BASED ON CITATION IDENTITY

Liming Shao[1] • Junpeng Yuan • Duanyang Xu

[1] *shaolm2012@istic.ac.cn*
Institute of Scientific and Technical Information of China, Beijing 100038, China

Introduction

This article provides an alternative perspective for measuring impact of an author by studying the set of all authors who cite the author. Besides, we propose a new algorithm which gives different weights for the authors in the cited author list. Based on this algorithm, a new indicator (citation identity degree) is proposed which can quantify the impact of each author. We test the indicator by evaluating author impact in Chinese information science community and compare this indicator with citation.

Data

The "Chinese Social Science Citation Index'' (CSSCI) was developed by Nanjing University. It is an important tool for inquiry and assessment of the major documentation in the area of humanities and social sciences.

Only information science fields included in CSSCI were identified for investigation. Data used for the study were limited to the period 2002–2011. Among all of the 9355 publications in information science field, there are 9335 publications that have authors and 8515 publications that have references. At last we chose 8505 publications which have both authors and references as our data.

Methods

We establish several new concepts: basic identity (BI) and citation identity degree (CID). We define basic identity as: BI = C / P, C are total citations with no self-citations and P are total papers. We also define citation identity degree as: CID=ΣBIi•Ni, BIi is the basic identity of author i and Ni are the times cited by author i. As is shown in figure1, Author A was only cited by author B, C, D and the cited times are N_b, N_c, N_d. The CID of author A is $CID_a=BI_b•N_b+BI_c•N_c+BI_d•N_d$. By doing this normalized weighted algorithm, we can quantify the influence of each author and compare the influence of different author.

Figure 1. Calculation method of citation identity degree

Results and discussion

At first, we have discussed the relationship between the number of this indicator and citations to see whether it can be used to evaluate the influence of authors. Then, we have analyzed the rank of citation identity degree.

Figure 2 shows the Q-Q plot of citation counts and citation identity degree values. In this graph, more nodes are distributed around the diagonal line, which indicate that citation identity degree values and citation counts have the same distribution. Because citation counts follow power-law distribution, the citation identity degree values also

follow power-law distribution. It means that citation identity degree, to a certain degree, also measures an author's academic impact.

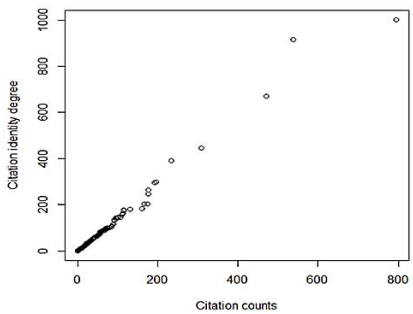


Figure 2. Citation counts and citation identity degree Q-Q plot

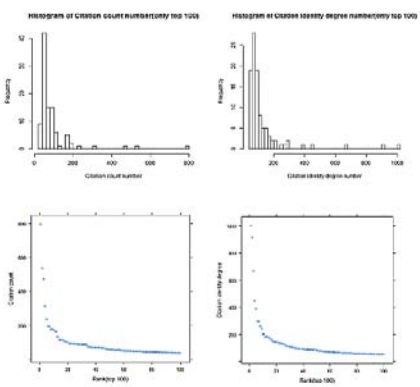


Figure 3. Distribution of citation and citation identity degree (only top 100)

Figure 3 shows the distribution of citation and citation identity degree (only top 100). Although citation counts and citation identity values have the same distribution, the citation identity degree is more evenly distributed. As the citation identity degree value is based on weighted algorithm, it gives the impact author more weight and the general author less weight, which lead to the discrepancy. Therefore, the citation identity degree values have a higher degree of distinction.

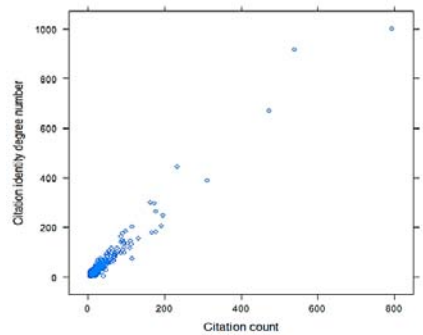


Figure 4. Citation counts and citation identity degree scores

Table 1. Top 20 authors in the citation identity degree and top 20 authors in the citation

No.	Author	CID	Author	C
1	Qiu Junping	1001.3	Qiu Junping	795
2	Bao Changhuo	915.75	Bao Changhuo	539
3	Ma Feicheng	669.17	Ma Feicheng	473
4	Yan Yimin	445.18	Wang Zhijin	311
5	Wan Zhijin	390.08	Yan Yimin	234
6	Wang Congde	299.25	Zhang Xiaoling	196
7	Liang Zhanping	296.76	Su Xinning	192
8	Lai Maosheng	262.94	Hu Changping	177
9	Zhang Xiaolin	245.63	Lai Maosheng	177
10	Su Xinning	204.50	Liang Zhanping	174
11	Liu Zhihui	203.46	Zhang Qiyu	166
12	Lu Taihong	183.69	Wang Congde	163
13	Hu Changping	181.40	Cheng Feng	132
14	Zhang Qiyu	177.67	Liu Zhihui	115
15	Yue Jianbo	175.30	Wu Xiaowei	115
16	Miu Qihao	163.66	Qin Tiehui	114
17	Cheng Feng	156.41	Pang Jingan	111
18	Meng Guangjun	147.16	Peng Jinli	108
19	Pang Jingan	145.41	Wen Youkui	102
20	Zhou Xiaoying	144.23	Lu Taihong	98

Figure 4 shows the scatter plot of citation counts and citation identity degree values. In this graph, more nodes are distributed around the diagonal line,

which indicates that the rank of citation identity degree value is similar to the rank of citation counts. We have calculated the coefficient (0.95), residual standard error (10.3) and p-value (<2.2e-16) of citation counts and citation identity degree values. We find that citation identity degree values and citation counts have high correlations for the lower levels, which indicates that the citation identity degree may yield useful values for the majority of authors. For higher values, the relatively low citation counts have a relatively high citation identity degree values. This result can be interpreted by some authors have a high impact but have rarely publications.

From table 1, we can easily conclude that the ranking order of citation identity degree and citation counts are mainly consistent, but the gap amongst the top 20 authors in citation identity list are far below that of citation counts list merely. Therefore, employing citation identity degree rank in authors' impact analysis possesses better distinction. Difference exists in citation identity degree ranking and citation counts ranking, i.e. Wang Chongde's rank in citation identity degree ranking has been raised 6 places; this is comparatively in line with Pro. Wang's identity as one of the earliest experts studying on teaching and researching the theories and methods of the information science and technology in China. Yue Jianbo, Miu Qihao, Meng Guangjun,and Chou Xiaoying's name are only appeared in top 20 citation identity degree ranking, take Pro. Miu for example, it known to all that Miu and Bao are equally famous in Chinese competitive information domain. Bao places second both in the list of citation identity degree rank and citation counts rank. Thus, Bao deserves taking a position in the top 20 of citation identity degree list.

However, there are also some limitations of this study. On the one hand, data collected by CSSCI have not included all articles that an author had published. On the other hand, we compared the new indicator only with citations instead of other indications.

Conclusion

The current study provides an alternative perspective for measuring author impact by learning the ideas of citation identity. We evaluate an author's impact by discussing the set of all authors who cite the author and propose a weighted algorithm. A novel indicator is also proposed. We test this indicator by evaluating author impact in Chinese information science community and compare this indicator with citation. Findings show that this new indicator provides a meaningful extension to the traditionally used citation counts for authors.

References

Ding, Y. (2011). Applying Weighted PageRank to Author Citation Networks. Journal of the American Society for Information Science and Technology, 62(2), 236-245.

Hirsch, J. E. (2005). An index to quantify an individual's scientific research output. Proceedings of the National Academy of Sciences of the United States of America, 102(46), 16569-16572.

White, H. D. (2001). Authors as citers over time. Journal of the American Society for Information Science and Technology, 52(2), 87-108.

DO NEW SCIENTISTS PREFER COLLABORATING WITH OLD SCIENTISTS? AND VICE VERSA?

Zhigang Hu[1] Haiyan Hou[2] Jianhua Hou[3] and Chunlin Jiang[4]

[1] *huzhigang@mail.dlut.edu.cn*
WISE Lab, Dalian University of Technology, Dalian (China)
Joint-Institute for the Study of Knowledge Visualization and Scientific Discovery, Dalian University of Technology(China)-Drexel University(USA), Dalian(China) and Philadelphia(USA)

[2] *seabirdofsummer@gmail.com*
WISE Lab, Dalian University of Technology, Dalian (China)

[3] *hqzhixing@gmail.com*
School of Humanities, Dalian University, Dalian (China)

[4] *chunlinj7873@163.com*
WISE Lab, Dalian University of Technology, Dalian (China)

Introduction

It is seemingly reasonable to consider that new and fledgling scientists might prefer collaborating with old and experienced scientists because they could benefit a lot from that, such as the satisfactory experiment condition, the brilliant ideas, and the opportunity to publish paper in top journals, etc. Moreover, collaboration of this type is generally considered as the way for the olds to guide the news. Thus, collaboration between new scientists and old ones is also called mentoring collaboration sometimes. Since mentoring collaboration is seemingly reasonable and valuable, we got a question: is it really happening?

Methods

In previous researches, Liang et al(2001) divided scientists into three groups: younger scientists (age<37), middle-aged (37<age<50), and elder (age>50), and explored how the three groups collaborate with each other. They found that Younger-Elder is the main type of age structures of scientific collaborations in computer science in China.

In this study, we designed a novel measurement to identify new and old scientists, which is called the academic age of scientists. When scientists publish their first publication in a field, their academic career is beginning. Thus we identified a scientist is new when his/her academic age is small and a scientist is old when his/her academic age is large. In this way, we could divide scientists into the new-scientist group and the old-scientists group simply and clearly.

Results

As the case, we chose all the authors who published papers in 1990 in the journals of four subject categories (JCR), namely Computer Science, Mathematics, Organic Chemistry, and Virology.

In 1990, the distributions of total scientists by academic age in the four subjects are shown in Figure 1.

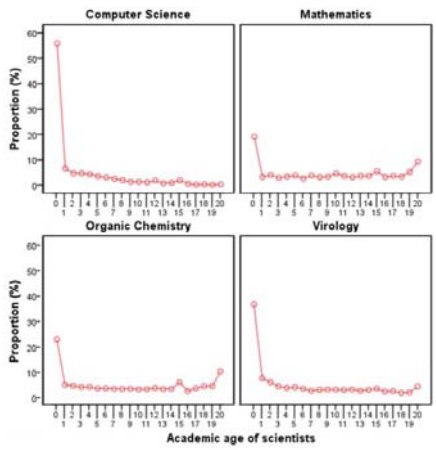


Figure 1. The distributions of the total scientists by academic age in the four subjects

Accordingly, we divided the authors into two equal group based on their academic ages in 1990. The younger group is named the new-scientist group, and the old one named the old-scientist group.

Do the new scientists prefer collaborating with the old ones?

We assumed that new scientists choose collaborators randomly and impartially, so the distribution of the new scientists' collaborators by academic age should be consistent with the distribution of all of the scientists by academic age as shown in Figure 1. Thus, by investigating the consistency between the expected distribution (as shown in Figure 1) and the actual distribution (as shown in Figure 2), we can give answer to above questions.

As shown in Figure 2, the two distributions are extremely fit with each other. However, in the experimental subjects, the proportion of collaborators aged 0 is obviously less than expected. The result means to some extent, new scientists prefer not to collaborate with the scientists which are also new. This is especially true in experimental subjects because new scientists require the laboratory and other scientific resources supported by their old collaborators.

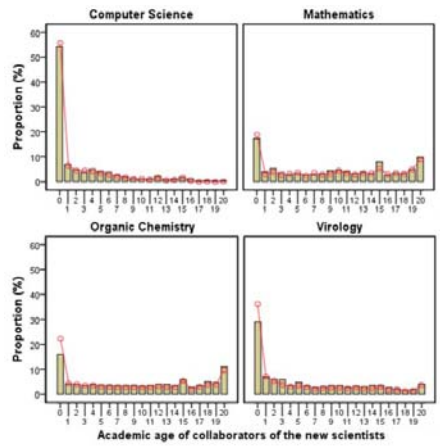


Figure 2. The distribution of the new scientists' collaborators by academic age

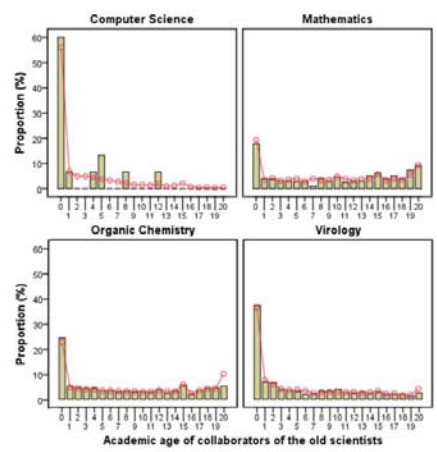


Figure 3. The distribution of the old scientists' collaborators by academic age

Do the old scientists prefer collaborating with the new ones?

Similarly, we examined how the old scientists choose collaborators and whether they are biased when conducting collaboration. Figure 3 shows the distribution of the old scientists' collaborators and the contrast to the expected distributions.

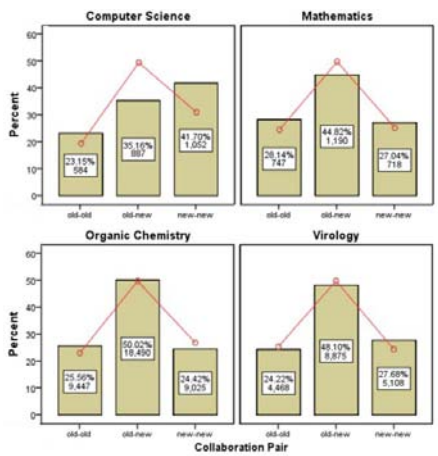


Figure 4 The distribution of the three kinds of collaboration pairs: old-old, old-new, new-new

Do scientists prefer mentoring collaboration or peer collaboration?

Further, we verified the above conclusions by investigating the probabilities of collaboration pairs of three different kinds, namely collaborations between two old scientists (old-old), collaboration between two new scientists (new-new), and collaboration between an old one and a new one (old-new). The former two are also called peer collaboration, and old-new collaborations are called mentoring collaboration.

Figure 4 shows the proportion of three kinds of collaborations. By comparing with the expected proportion (around 1:2:1), it is found that for theory subjects like Computer science and Mathematics, peer collaborations are preferable rather than mentoring collaborations.

Conclusion

In summary, scientists don't choose collaborators according to their academic age. For both new and old scientists, the distributions of their collaborators by academic age are consistent with their expected distributions.

Only when examining this issue more deeply, we can find some inclinations about collaborator choosing. For example, in theoretical subjects, peer collaborations are more desirable than mentoring collaborations; while in experimental subjects, the opposite inclination appears.

References

Liang L, Kretschmer H, Guo Y, Beaver DD. Age structures of scientific collaboration in Chinese computer science. *Scientometrics*. 2001;52(3):471–86.

DO SMALL AND MEDIUM SIZED BUSINESSES CLAIM FOR SMALL ENTITY STATUS? THE CASE OF MIT AND STANFORD UNIVERSITY SPINOFFS

Ahmad Barirani

ahmad.barirani@polymtl.ca
Polytechnique Montréal, P.O. Box. 6079, Downtown Office, Montreal, Qc, H3C 3A7, Canada

Introduction

The USPTO gives the opportunity for persons, non-profit organizations and businesses with less than 500 employees to claim for the small entity status (the Status) which entitles them to pay lower (50% discount) patent maintenance fees. Studies have used the Status to distinguish between patents granted to large and small firms (Allison and Lemley, 2000, 2006; Park and Park, 2006; Allison et al., 2009; Fernandez-Ribas, 2010; Bessen, 2008; Alcacer et al., 2009).

These studies make the implicit assumption that only large corporations do not claim for the Status. However, there could be strategic reasons for small and medium sized enterprises (SMEs) not to apply for the Status. Among the requirements for qualification, it is stipulated that at the time that the Status is claimed, there must be no obligation to assign, grant, convey, or license any rights to the invention to any entity that would not, in turn, qualify for small entity status (Patent and Office, 2001, § 1.27(a)(2)(i)). Thus, SMEs that are going to license their patents to large firms will not be able to claim for the Status. To what extent are these cases where SMEs don't apply for small entity status widespread?

Methodology

A sample of university spinoffs that originate from the MIT and Stanford University are employed. After cleaning for strings that describe the firm's legal entity (such as Inc., Corp., etc.), the set of patents granted by the USPTO to these spinoffs is extracted. From this set of extracted patents, only those that are assigned to firms that are located in the same state as their university of origin (i.e. Massachusetts for the MIT and California for Stanford University) are considered. This will avoid taking into consideration homonyms. From this set of firms for which patents have been matched in the USPTO, search on LinkedIn website is performed in order to gather information about current number of employees. Since not everyone has a LinkedIn account, this information is not always accurate. Therefore, only those firms which have currently less than 350 employees will be used for further analysis. This cushion of 150 employees appears to be reasonable taking into account that, for a total of 155 million workers in the US workforce, more than 74 million LinkedIn accounts have been created in the US and that the use of the service is more widespread among hi-tech professionals.

Patents owned by these firms are then linked to USPTO's maintenance fee events database, from which those that are related to the Status are identified. Patents that are not associated with the Status will represent *false large entity patents*.

Results

To date, the MIT and Stanford University have produced 131 and 227 spinoffs respectively. From these 358 firms, 105 were found to have patents in the USPTO database. Based on searches made on LinkedIn, 60 of these firms are active today and have less than 150 employees. Table 1 shows the number of patents granted to each of these firms, and the number of patents for which a small entity event is found in the maintenance fee database. As we can see, a substantial number of patents (more than 30%) produced by these SMEs are not claimed as small entities.

Table 1. Maintenance fee events for active spinoffs from the MIT and Stanford University.

Firm	All patents	Small entity patents
anacor pharmaceuticals	4	2
applied genomics	2	1
arbor vita	11	6
art technology group	10	7
avantec vascular	15	1
barcelona design	12	0
biocardia	24	24
brion technologies	29	15
caveo technology	1	1
cellgate	12	9
comentis	2	0
cooligy	29	16
corcept therapeutics	7	6
corgentech	2	2
coverity	5	4
e ink	143	136
ember	4	4
fluidigm	25	22

Table 1. Maintenance fee events for active spinoffs from the MIT and Stanford University (continued).

Firm	All patents	Small entity patents
general mems	1	0
harmonix music systems	8	4
kosan biosciences	91	50
lightbit	3	3
lightconnect	9	8
lyncean technologies	5	5
microbar	6	5
molecular nanosystems	2	2
nearlife	4	4
neophotonics	19	1
neurocrine biosciences	69	31
novariant	29	28
open ratings	2	1
optimedica	2	2
panorama research	1	1
picarro	20	16
pixim	46	46
predicant biosciences	5	1
rigel pharmaceuticals	140	114
sabio labs	2	0
sensable technologies	25	25
senvid	3	2
sirf technology	160	89
spinal modulation	4	4
spiracur	1	0
stemcells	2	2
t-ram	96	94
tagsense	1	1
telik	27	2
terastor	45	7
thingmagic	6	0
tosk	7	5
trellis bioscience	7	6
via pharmaceuticals	3	1
viisage technology	5	4
voltage security	14	14
way systems	3	3
xgene	2	2
zomed international	8	8
zyomyx	21	19
Total	***1242***	***867***

Conclusions

Studying a sample of spinoffs originating from the MIT and Stanford

University, it can be found that a substantial part of patents produced by these firms where not claimed under the small entity status. It would thus appear that the small entity status is not a great source of information for finding whether a firm is large or not. However, given the strict rules under which one can claim for the small entity status, it is safe to conclude that a sample of patents for which events in the maintenance fee database are associated with the small entity status have been indeed assigned to a non-profit organization, an individual or an SME.

An inherent limitation to this research resides in the fact that a small sample of university spinoffs is used. Another limitation consists in using LinkedIn as a source of information about the current size of a company. Data obtained from LinkedIn is not always reliable as it results from the entry of information by the users of the service.

The discussion in this study can be complemented by performing descriptive and inferential statistics on different variables (such as firm age or industry) in order to have a better indication of how false large entity patents are distributed among SMEs. Trend analysis can also be interesting in order to observe whether there is a tendency for SMEs to claim less often for the Status.

References

Alcacer, J., Gittelman, M., and Sampat, B. (2009). Applicant and examiner citations in U.S. patents: An overview and analysis. *Research Policy*, 38(2):415–427.

Allison, J. and Lemley, M. (2000). Who's Patenting What-An Empirical Exploration of Patent Prosecution. *Vanderbilt Law Review*, 53:2099.

Allison, J. and Lemley, M. (2006). (Unnoticed) Demise of the Doctrine of Equivalents, *The Stanford Law Review*, 59:955.

Allison, J., Lemley, M., and Walker, J. (2009). Extreme Value or Trolls on Top? The Characteristics of the Most Litigated Patents. *University of Pennsylvania Law Review*, Vol. 158, No. 1, December 2009, Stanford Public Law Working Paper No. 1407796.

Bessen, J. (2008). The value of U.S. patents by owner and patent characteristics. *Research Policy*, 37(5):932–945.

Erickson, G. (1996). Environment and innovation: The case of the small entity. *Industrial Marketing Management*, 25(6):577–587.

Fernandez-Ribas, A. (2010). International patent strategies of small and large firms: An empirical study of nanotechnology. *Review of Policy Research*, 27(4):457–473.

Park, G. and Park, Y. (2006). On the measurement of patent stock as knowledge indicators. *Technological Forecasting and Social Change*, 73(7):793 – 812.

Patent, U. S. and Office, T. (2001). *Manual of Patent Examining Procedure*.

DOES SCIENTIFIC KNOWLEDGE PLAY A ROLE IN PUBLIC POLICIES? A CONTRIBUTION OF SCIENTOMETRICS TO POLITICAL SCIENCE: THE CASE OF HTA.

Cyril Benoit[1] and Philippe Gorry[2]

[1] *cyril.benoit@scpobx.fr*
Centre Emile Durkheim UMR CNRS 5116, Sciences Po Bordeaux, 11, Allée Ausone- – 33607 Pessac (France)

[2] *philippe.gorry@u-bordeaux4.fr*
Research Unit in Theoretical and Applied Economics, UMR CNRS 5113, University of Montesquieu Bordeaux IV, 33607 Pessac (France)

Introduction

The social use of scientific knowledge can take various forms. In the case of public policies, most researches insist on the capacity of decision-makers to technicize issues with scientific-based knowledge. In the health sector, previous work was focused on the link between the field of academic research and that of decision-making (Lavis et al., 2006). Recently, some demonstrated the appropriateness of bibliometric studies in observing the factors that influence the use of scientific-based knowledge in public policies (Macias-Chapula, 2012). According to them, we propose to enhance this approach in health public policies analysis. Indeed, many of them are based on wide process of knowledge circulations.

To illustrate this argument, we propose to lead a scientometric analysis of conditions of emergence of a public action instrument (Lascoumes et al., 2007): Health Technology Assessment (HTA). It has been defined as a form of policy research that examines short-and long-term consequences of the application of health technologies (Banta, 2009). HTA aims to link regulation of the healthcare system, quality of care, and payment for care (Banta et al., 2010). Since the mid-2000s, the whole of OECD countries and most of middle-income countries have incorporated various HTA tools to their decision-making process. They were generally institutionalized as "HTA agencies"; which provide the stakeholders scientific-based assessment of cost and benefits of a wide set of medical devices. Its scientific dimension can be considered as the main feature of HTA: assessments are led by professional researchers, proposing a rational aid to external decision-makers, strongly related to dedicated research centre.

Nevertheless, we believe that this phenomenon raises three major policy issues: what were the scientific conditions for HTA success? Who were the individuals, and why they linked academic research and decision-making process? Which are the factors for a virtuous institutionalization of HTA in a given country?

Purpose

The purpose of this work was to undertake a historic, descriptive and quantitative comparative analysis of scientific conditions of emergence of HTA as a *public action instrument* applying scientometrics to political science questions.

Methods

A literature search on HTA scientific production was conducted both in PubMed & Scopus databases Then, we led a bibliometric, network and descriptive statistics analysis on the corpus (2081 publications) with various analysis tools (KNALIJ, Matheoanalyzer, Intellixir). The landscape of the HTA field so obtained was deployed in 3 comparative registers. A micro-level approach through a network analysis was led in order to identify HTA main experts with their pattern of collaborations. Then, we replaced it in a macro-level perspective in mapping the worldwide academic research. Finally, we explored the link between these scientific dimensions and healthcare system with PCA analysis.

Results

The temporal distribution of the publications suggests a peripheral status of our concept in the literature, until a turning point in the late 1980's. Then the growth of the publications became exponential with an acceleration rate superior to related disciplines s (Fig. 1).
HTA publications were produced by authors grouped in some collective research schemes organised around one 'major author' (red coloured on Fig.2). Indeed, the distribution of publications by authors follow a Lotka's Law.
A map of concepts was draw to identify the main scientific streams (Fig.3), and time-series analysis help us to corroborate with HTA institutionalization.

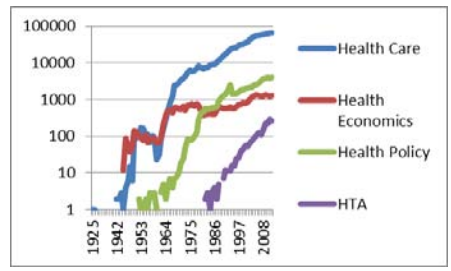


Figure1. Number of HTA publications per year in comparison to related disciplines

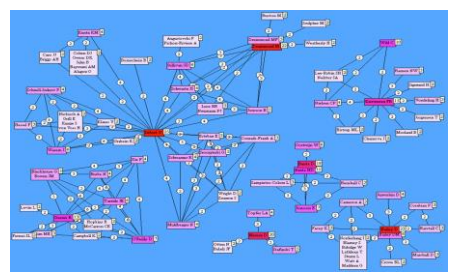


Figure 2. Network analysis of HTA scientific collaborations

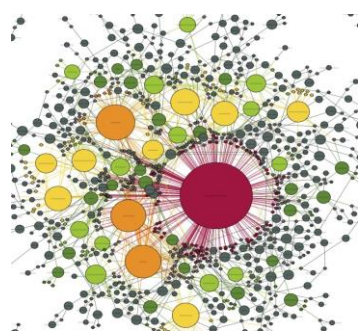


Figure 3. Network map of HTA main concepts

Country HTA publications productivity (Fig.4) is matched in comparison with health expenses & gross product in several countries (strongest ratio: at the top, right; weakest ratio: at the bottom, left) (Fig. 5).

Figure 4. Choropleth map of HTA publications

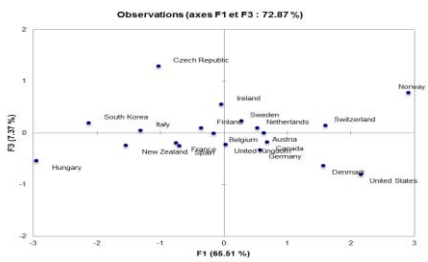


Figure 5. PCA plot of country HTA research according to GDP & health expenses.

Discussion

For the case of HTA, it appears that a small scholar's community have gradually become independent in the scientific field. Although it seems to be closely linked to public health policies, the scholars working on HTA are not submitted to public purchasing. The growth of HTA as a field autonomous from decision-making field and other disciplinary fields permit to consider it as a scientific discipline, in the sense of P. Bourdieu (Albert et al., 2011). In Fig. 4 & 5, the absence of formal relation between the importance of HTA publications and an early institutionalization of the concept confirm this view. Its career in the different fields must be distinguished. Despite this autonomy, scholars are in position to develop concept that can come out to legitimate new orientations in public policies (Roger, 2010).

When scholars seem to influence decision-making process, the position of each of them in their field must be reconstructed. A bibliometric analysis appeared as an appropriate tool to conduct such investigation in political science.

References

Albert M. & Kleinman D.L (2011). Bringing Pierre Bourdieu to Science and Technology Studies. *Minerva*, Vol. 49, 263-273.

Banta, D. (2009).What is technology assessment? *International Journal of Technology Assessment in Health Care*, Vol. 25, 7-9.

Banta, D., Mathijssen, J., & Oortwjn, W., (2010). The role of health technology assessment on pharmaceutical reimbursement in selected middle-income countries. *Health Policy*, Vol. 95, 174-184.

Lascoumes, P. & LeGalès, P. (2007). "Understanding public policy through its instruments", *Governance*, Vol. 20, 1-21.

Lavis, J. N., Lomas, J., Hamid, M., & Sewankambo, N. K. (2006). Assessing country-level efforts to link research to action. *Bulletin of the World Health Organization*, n°84, 620-628.

Macias-Chapula, C. (2012). Comparative analysis of *health public policy* research results among Mexico, Chile and Argentina, *Scientometrics*, 1-14.

Roger, A. (2010). Scholarly constructs and the legitimization of European policies. *Revue Française de Science Politique* (English), Vol. 60, 1-22.

THE EARLIEST PRIORITY SELECTOR FOR COMPILING PATENT INDICATORS

Douglas H. Milanez[1], Mateus G. Milanez[2], Leandro I. L. de Faria[3], Roniberto M. do Amaral[4] and Jose A. R. Gregolin[5]

[1]*douglas@nit.ufscar.br*
Federal University of Sao Carlos, Washington Luis Highway, km 235, São Carlos – SP (Brazil)

[2]*milaneza@gmail.com*
MLNZ IT Consultancy LTDA, Saverio Talarico Avenue, 220, São Carlos – SP (Brazil)

[3]*leandro@nit.ufscar.br*
Federal University of Sao Carlos, Washington Luis Highway, km 235, São Carlos – SP (Brazil)

[4]*roniberto@nit.ufscar.br*
Federal University of Sao Carlos, Washington Luis Highway, km 235, São Carlos – SP (Brazil)

[5]*gregolin@nit.ufscar.br*
Federal University of Sao Carlos, Washington Luis Highway, km 235, São Carlos – SP (Brazil)

Introduction

Patent indicators provide an effective opportunity to depict patterns of technological development from countries and competitors. The OECD Patent Statistics Manual (2009) provides basic criteria for compiling patent indicators and two important methodological choices are the reference date and the country of attribution. Due to the legal constraints and administrative delays, the manual recommends using the date from the first priority as it is closest to the date of the invention. Concerning the country of origin, patents can be assigned on the basis of the priority country or addresses of the inventor or the patentee. Although the suggestion is to use the inventor's country in order to reflect the country's inventive activity (OECD, 2009), this information is not always available on databases, as in the case of Derwent Innovations Index (DII), whose patent family records include only the priority information as a source of country of origin. A patent family can be defined as a group of patent publications on a single invention (Simmons, 2009). Usually, a patent family claim the same priority (first filing), but they can also have multiple priorities as a result of the variations in legal regulation among countries or rules for creating patent families (DII, 2013; Simmons, 2009). According to DII (2013), only around 2% of all patent applications indexed claim multiple priorities. However, multiple priorities might affect the analysis of patent indicators as different countries and dates can be wrongly included in the final indicator. This

paper aims to present software that selects the earliest priority from patent families indexed in the DII database, providing the correct information to compile patent indicators. A comparison of indicators using all priorities or just the earliest one was also conducted.

Materials and Methods

Patent data sample and analysis

A test dataset of nanotechnology patents was retrieved from the DII database using the modularized Boolean search strategy suggested by Porter et al. (2008). The search was carried out on 23 January, 2013 and 189,481 patent family records were collected considering the time spam from 1995 to 2012. After treating records in the Earliest Priority Selector (EPS) software, the data was imported to the bibliometric software Vantage Point (version 7.0, Search Technology Inc, US). The accurate number of priorities was counted for each record in order to quantify the percentage of multiple priority records. The influence of using all priorities or the earliest was carefully examined for the top 15 countries and for the years.

The Earliest Priority Selector

The EPS[180] is software developed in Python language that aims to choose the first priority of records from DII with multiple priorities. Python is an open source license programming language with clear syntax, extensive standard libraries and modules for a range of task (Python, 2013). The EPS imports a single file with all records collected from database and its algorithm checks the priority field (called PI) as described below:

If the record presents just one priority, the EPS logically keeps it as the earliest priority;

If the record presents two or more priorities, the EPS compares the priority dates and chooses the earliest.

Although most of the records were accurately treated with the described algorithm, a few records presented different earliest priority numbers with the same date, as exemplified in Table 1. In all of these cases, however, a WO priority number was among the earliest priorities. Thus, the EPS algorithm picks up the WO priority as the earliest date. In addition, the earliest priority selected by EPS was added into a new field (called PO), which is also included in Table 1.

Table 1. Example of multiple priorities from DII records (WO2004001100-A1).

Field	Priority number	Priority date
PI	CN80158992	22 Jul 2009
	JP523510	22 Jul 2009
	WOJP063106	22 Jul 2009
	US353569	19 Jan 2012
	JP093340	16 Apr 2012
PO	WOJP063106	22 Jul 2009

Results and discussion

From the 189,481 nanotechnology patent families, 75.9% showed just one priority, and then no treatment was necessary. By contrast, from the 24.1% with multiple priorities, 17.0% were records with two priorities and 3.91% with three priorities. Just a few patent families fit in the special case shown in Table 1, and they represent 0.55% from the whole nanotechnology patent sample and 2.28% from the multiple priorities dataset. In this case, even if the selected country is not the correct one, it is clear that the error will not affect any ranking or analysis. Table 2 shows the ranking

[180] The software is freely available at http://www.nit.ufscar.br/index.php/software/154-earliest

of the top 15 countries considering the information from all priorities (PI field) and from the earliest priority selected by EPS (PO field).

Table 2. Differences in country ranking

Country Code	PI Field		PO Field		Change (%)
	P	NP	P	NP	
CN	1	54293	1	53224	-1.97
US	2	53329	2	50837	-4.67
JP	3	33015	3	32555	-1.39
KR	4	22553	4	19371	-14.1
DE	5	8384	5	8115	-3.21
CA	6	4712	17	321	-93.2
TW	7	4492	6	4427	-1.45
FR	8	3530	8	3507	-0.65
RU	9	3461	9	3414	-1.36
GB	10	2596	10	2480	-4.47
IN	11	1244	11	1062	-14.6
AU	12	1237	13	557	-55.0
IT	13	730	12	727	-0.41
ES	14	544	14	543	-0.18
BR	15	491	16	431	-12.2

Generally, no significant differences were observed in the country´s position (P) neither great changes in the number of patents (NP), except for Canada, Korea Australia, India and Brazil. In the case of Canada, it is known that many Canadian firms file patents first in the USA, followed by a possible extension in Canada at a later stage of the process (OECD, 2009). Therefore, the first priority was related to the USA not Canada. Although the change of the NP was high for Korea Australia, India and Brazil, their position in the ranking remained pretty much the same. A percentage decrease in the annual number of patents was observed when the year of the earliest priority was taking into account instead of all years from the priority field. On average, the number of patent families dropped 17.7% ± 1.71% for each year in the period analyzed.

Conclusion

No significant changes in the country´s performance were observed when the earliest priority was applied instead of all priorities. The fall of patent numbers is similar for all years from the period 1995-2010 and the use of the earliest priority is a better option due to the proximity to the invention, as recommended by the OCDE Patent Manual (2009). Although there is a bias of earliest priority with different numbers and equal dates (Table 1), the solution provided by the EPS does not significantly affect the final indicators. The accurate selection of the earliest priority also helps patent databases that do not provide the country of origin of the applicants or the inventors, as in the case of DII.

Acknowledgements

The authors are grateful to the Brazilian National Council for Technological and Scientific Development (process number 160087/2011-2), the São Paulo Research Foundation (process number 2012/16573-7) and the Graduate Program in Materials Science and Engineering at the Federal University of São Carlos.

References

Derwent Innovations Index. (2013). What is a Patent Family? Retrieved February 26, 2013 from http://images.webofknowledge.com/WOKRS59B4_2/help/DII/hs_book_part3.html

OECD. (2009). *OECD Patent Statistics Manual*. Retrieved February 26, 2013 from http://dx.doi.org/10.1787/9789264056442-en

Porter, A. L., Youtie, J., Shapira, P., & Schoeneck, D. J. (2008). Refining search terms for nanotechnology. *Journal of Nanoparticle Research*, *10*(5), 715–728.

Python. (2013). *About Python*. Retrieved February 26, 2013 from http://www.python.org/about/

Simmons, E. S. (2009). “Black sheep” in the patent family. *World Patent Information*, *31*(1),11–18.

EFFICIENCIES IN NATIONAL SCIENTIFIC PRODUCTIVITY WITH RESPECT TO MANPOWER AND FUNDING IN SCIENCE

Aparna Basu

aparnabasu.dr@gmail.com
CSIR-National Institute of Science Technology and Development Studies,
Dr. K.S. Krishnan Marg, Pusa Gate, New Delhi 110012, (India)

Introduction

Due to recession in the world economy there is a general trend towards a reduction in growth of R&D expenditure in the G7 countries. Asian countries like China and Korea have significantly increased their share of investment in R&D. China has also substantially increased its output of scientific papers. It is now second to the United States in terms of both papers and manpower. We ask how increased investment of resources translates into outputs. Do developed countries make more *efficient* use of their resources? We use the notion of efficiency of scientific output for inputs like manpower and investment at the country level to answer these questions.

Albuquerque's Model

Albuquerque (2005) proposed a simple model that linked output indicators to development. He showed that paper production and patent production ratio changes as countries become more developed. He termed this technological 'maturity' and took the ratio of papers to patents, normalized by population, as *efficiency*. (The ratio decreases as a function of development.) In Albuquerque's model, less developed countries fall on a line separated by a threshold from developed countries when patents are plotted against papers (Fig.1).

Data and methodology

Data on scientific papers is taken from the SCI-Expanded and USPTO for the years 2007 and 2008. GDP and GERD are both adjusted to Purchasing Power Parity (PPP) and compared for selected countries for the years 2002 and 2007. Manpower is measured in terms of Full Time Equivalents (FTEs) engaged in R&D. Data on GERD and GDP are obtained from the UNESCO Science Report 2010.

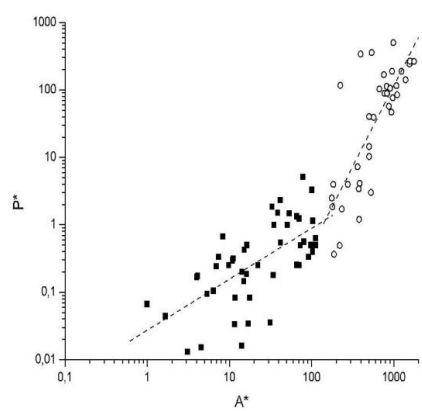


Fig.1. Papers A and Patents P (per mill. inhabitants) plotted for less developed countries (squares) and developed countries (circles). From Albuquerque 2005.

Our definition of *efficiency* of the science system is the output of papers and patents as a ratio of inputs, manpower in R&D and expenditure. This must be distinguished from Albuquerque's definition. Our definition more closely captures the efficiency with respect to actual expenditure and manpower. We have two dimensions for the outputs, patents and papers, and two dimensions for inputs, research expenditure and manpower, leading to a total of 4 indicators of efficiency.

Table 1. GERD and GDP shares of selected countries

Country	GDP share 2002	(2) GDP share 2007	GERD share 2002	GERD share 2007	GERD share/ GDP share 2002	GERD share/ GDP share 2007	GERD/ GDP 2007-2002
EU	25.3	22.5	26.1	23.1	*1.0*	*1.0*	*0.0*
USA	22.5	20.7	35.1	32.6	*1.6*	*1.6*	*0.1*
China	7.9	10.7	5	8.9	*0.6*	*0.8*	*0.2*
Japan	7.4	6.5	13.7	12.9	*1.9*	*2.0*	*0.1*
Germany	4.9	4.3	7.2	6.3	*1.5*	*1.5*	*0.0*
India	3.8	4.7	1.6	2.2	*0.4*	*0.5*	*0.1*
France	3.7	3.1	3.9	3.4	*1.1*	*1.1*	*0.0*
UK	3.7	3.2	3.9	3.4	*1.1*	*1.1*	*0.0*
Italy	3.3	2.8	2.2	1.9	*0.7*	*0.7*	*0.0*
Brazil	2.9	2.8	1.6	1.8	*0.6*	*0.6*	*0.1*
Russia	2.8	3.2	2.0	2.0	*0.7*	*0.6*	*-0.1*
Mexico	2.1	2.3	0.5	0.5	*0.2*	*0.2*	*-0.0*
Korea	2	1.9	2.8	3.6	*1.4*	*1.9*	*0.5*
Canada	2	1.9	2.4	2.1	*1.2*	*1.1*	*-0.1*
Australia	1.3	1.2	1.3	1.4	*1.0*	*1.2*	*0.2*

Efficiency is defined as,

Expenditure efficiency *EE(Pap)*
= Papers/GERD (1)

Manpower efficiency *ME(Pap)*
= Papers/Manpower (2)

For patents,

Patent Expenditure Efficiency *EE(Pat)*
= Patents/GERD (3)

Patent Manpower Efficiency *ME(Pat)*
= Patents/Manpower (4)

Table 1 below shows the base data of GERD and GDP, and their ratios.
Developed countries have higher GERD shares as compared to GDP shares (*GDP is the Gross Domestic Product*), the Gross Expenditure on R&D (GERD) being taken as the expenditure on the creation of new knowledge (Hollanders and Soete, 2010). Such countries with GERD/GDP >1 are Japan, USA, Germany, UK, France, Korea, Australia.
Table 2 below shows inputs and outputs to the science system.

Table 2Manpower, GERD, Papers and Patents for selected countries

Country	*GERD Bn $ PPP*	*Manpower (FTE's)*	*Papers SCI 2007*	*Patents USPTO 2008*
Australia	15.36	87,140	28313	1516
Brazil	20.2	1,33,266	26482	124
Canada	23.96	1,39,011	43539	3806
China	102.4	14,23,380	104968	7362
France	42.89	2,15,755	57133	3631
Germany	72.24	2,90,853	76368	9713
India	24.79	1,54,827	36261	741
Italy	22.12	96,303	45273	1836
Japan	147.9	7,09,974	74618	33572
S. Korea	41.3	2,21,928	32781	6424
Mexico	55.9	37,930	8262	81
Russia	23.4	4,51,213	27083	286
Spain	19.34	1,30,896	35739	363
UK	41.04	2,61,406	71,302	4007
USA	398	14,25,550	2,72,879	81811

Analysis

In Fig.2 we plot the expenditure and manpower efficiency of papers. Italy has the highest efficiency *EE(Pap)* and *ME(Pap)* in the production of scientific papers. Japan, Korea, USA and Germany have values of expenditure efficiency below average. The unexpectedly low values of efficiency are surprising for Japan and the USA as these, along with Korea and Germany, have shown an increase in the GERD/GDP ratio (Table 1.)
In Fig.3 we plot the *EE(Pat)* against the Expenditure Efficiency of Papers

EE(Pap). This is the analogue of Albuquerque's model (Fig.1) with our definitions of efficiency. If we compare the curves in the two figures, we note that there is a distribution of points along the X-axis in Fig. 3 that corresponds to the distribution in Fig.1, with a few prominent outliers. For the four countries that stand away from this distribution, namely Japan, USA, Korea and Germany, higher patent efficiency is not linked to higher paper efficiency.

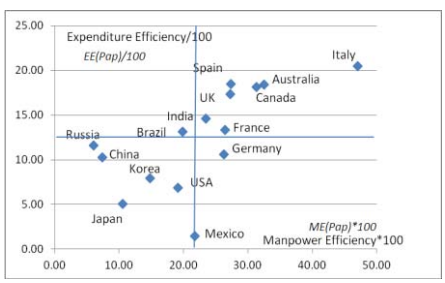


Fig 2. Efficiency in the production of papers w.r.t. Manpower and Expenditure

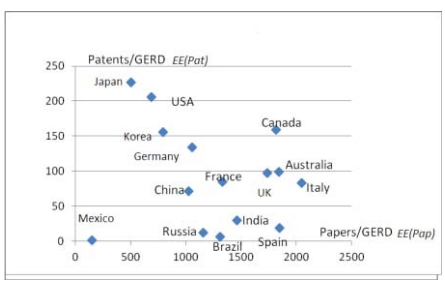


Fig. 3. Patent Efficiency with respect to GERD vs. Paper efficiency with respect to GERD

There appears to be an inverse relationship between patents and papers in a departure from Albuquerque's model.

A more detailed look at GERD figures for these countries shows 78.2% of total R&D expenditure in Japan came from the business sector. It was 67.3% in the US, 67.6% in Germany and 45.1% in the UK (2008 figures). These high values provide an explanation, viz., GERD funding from the business sector is driving patent production in favour of paper production.

Summary

Some findings are listed below. 1) Italy has the highest efficiency in the production of scientific papers, with respect to manpower and investments. 2) Japan has the highest efficiency in patent production. 3) Certain countries that have shown increased expenditures in R&D, showed at the same time a low efficiency in the output of scientific papers, which was unexpected. This appeared to be compensated in by high efficiency in the production of patents. We find that there is a shift from publications toward patenting apparently driven by increased investments from the business sector in countries like Japan, USA, Korea and Germany.4) It follows that Albuquerque's model with a simple demarcation between developed and developing countries is no longer true. Better definitions of *efficiency* are required. The trend in more developed countries is not to increase the output of papers, as suggested by the model, but to decrease the output of papers in favour of patents.

References

Albuquerque, E. (2005) Science and Technology systems in less developed countries. In H.Moed, W. Glanzel, U.Schmoch (Eds.) *Handbook of Quantitative Science and Technology Research* (pp759-778) Kluwer Academic Publishers.

Hollanders, H. & L. Soete (2010). The growing role of knowledge in the global economy. In *UNESCO Science Report 2010,* UNESCO Publishing.

UNESCO Science Report (2010) UNESCO Publishing.

EMERGENCE OF KEYWORDS IN WEB OF SCIENCE VS. WIKIPEDIA

Christine Rimmert[1] , Edith Rimmert[2] and Holger Schwechheimer[3]

[1] *christine.rimmert@uni-bielefeld.de*
Bielefeld University, Interdisciplinary Bibliometric Group, Universitätsstr. 25, 33615 Bielefeld (Germany)

[2] *edith.rimmert@uni-bielefeld.de*
Bielefeld University, Library, Universitätsstr. 25, 33615 Bielefeld (Germany)

[3] *holger.schwechheimer@uni-bielefeld.de*
Bielefeld University, Interdisciplinary Bibliometric Group, Universitätsstr. 25, 33615 Bielefeld (Germany)

Introduction

Topics appearing in scientific literature are also discussed in the open web - thus it is worth asking whether the open web is an interesting source for bibliometric issues. This paper focuses on keywords in the Web of Science (WoS) and associated Wikipedia pages. Are WoS keywords also represented in Wikipedia? When do they emerge in Wikipedia compared to their first appearance in the WoS?

Methods

Four samples of keywords (appeared after 2007) were taken from the WoS:

- random selection of 100 author keywords ('AU' in the following),
- the 100 most frequent AU keywords,
- random selection of 100 KeyWords Plus® ('PL' in the following) and
- the 100 most frequent PL keywords.

For each keyword in each sample the occurrence in Wikipedia was checked using four matching conditions:

1. exact match (the keyword is equal to a title of a Wikipedia page),
2. redirected match (searching the keyword leads to a redirection to another page),
3. chapter match (the keyword is equal to a chapter on a Wikipedia page) and
4. similar match (a Wikipedia page is named in a similar way),

where 'similar' was restricted to differences concerning singular-plural, well known abbreviations (e.g. 'DNA Seq'), word permutations, typing errors, part of speech and missing stop words in order to prevent errors. If the WoS keyword contains two keywords (e. g. 'keyword1 and keyword2'), an exact match of keyword1 or keyword2 is counted as an exact match. The same applies for keywords that differ only by a hyphen. The 'start' was defined as the year of the first appearance in the WoS, respectively the emergence of the Wikipedia page.

Results

Matches

Table 1 shows the match results. While the random selections have high percentages of keywords that could not be found in Wikipedia (AU:83%, PL:86%), the situation is better for top keywords (in particular for AU). Overall, however, many keywords could not be found (they often seem to be too specific or cannot be matched without specific expertise).

Table 1. Keyword occurrence in Wikipedia.

	1	2	3	4
Random AU	2	5	4	6
Random PL	9	0	2	3
Top 100 AU	27	20	4	16
Top 100 PL	15	6	3	5

Start years

In order to compare start years, exact, redirected and similar matches were taken from the top samples. Table 2 and Figure 1 show the differences (in case of only one dot, the start years are the same).

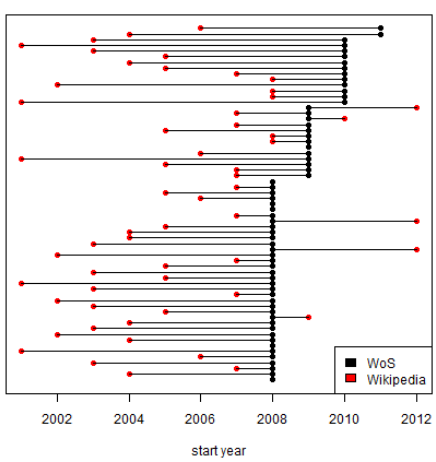


Figure 1. Comparison of Wikipedia and WoS start years (AU keywords).

In both samples the keywords seem to appear in Wikipedia before or simultaneously to their start in the WoS in the vast majority of cases (Table 2).

Table 2. Differences of start years.

	WoS<Wiki	WoS=Wiki	WoS>Wiki
PL	5	11	10
AU	5	6	52

But including the variants found while searching Wikipedia pages (single words, similar words and words leading to a redirection) and looking at them as keywords in the WoS again provides a different picture. 32 of the variants found in the AU sample appear themselves as WoS keywords; 25 (2 in the PL sample) of them have start years before 2001 – the launching year of Wikipedia – thus a comparison of start years is no longer appropriate.

Conclusion and further work

In the top samples more keywords could be found as titles of Wikipedia pages than in the random samples. It is possible that more matches could be found with specific expertise. Among the matched keywords, in the vast majority of cases the start in Wikipedia at first sight seems to be earlier than the start in the WoS. In order to receive reliable results, however, more matched keywords would be needed. For the future it would be interesting to look at other information of Wikipedia pages, e.g. redirections, links (as hints for keyword unification or field delineation tasks) and page view statistics, e.g. to ompare the growth of publications/citations in scientific databases to the growth of page views in Wikipedia (available since 2008), see Figure 3 for an example.

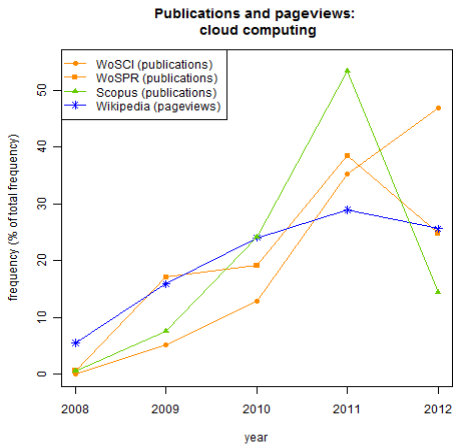


Figure 3. Example for a comparison of page views and publication counts.

Acknowledgement

Certain data included herein is derived from the "Science Citation Index Expanded (SCIE) - Tagged Data" prepared by Thomson Reuters (Scientific) Inc. (TR®), Philadelphia, Pennsylvania, USA: © Copyright Thomson Reuters (Scientific) 2012. All rights reserved.

References

Wikipedia.[http://en.wikipedia.org/]

ENTROPY-BASED DISCIPLINARITY INDICATOR: ROLE TAXONOMY OF JOURNALS IN SCIENTIFIC COMMUNICATION SYSTEMS.

Jorge Mañana-Rodríguez[1]

[1] *jorge.mannana@cchs.csic.es*
Consejo Superior de Investigaciones Científicas, Centro de Ciencias Humanas y Sociales. Madrid, España, C/Albasanz 26-28, P.O. Box 28037. T.N.: (0034) 916022795.

Introduction

In recent decades, the study of multidisciplinarity / disciplinarity has emerged as a core topic in science and technology studies and information & library science. IDR, Inter Disciplinary Research (Wagner et al. 2011), is a key aspect both for policymakers and researchers (National Academies, 2005).

In this contribution, the author presents a new indicator for measuring the degree of disciplinarity or multidisciplinarity of scientific journals, both in the citing and cited dimensions. The specialization (or disciplinarity) degree of the indicator is directly proportional to the percentage of cites/references from/to other journals of the same discipline and inversely proportional to the Shannon entropy of the distribution of cites/references from/to journals in disciplines different from the journal one.

Objectives:

In this work the author seeks for a vector based indicator which values attempt to capture the degree of disciplinarity or multidisciplinarity (the latter corresponding to the definition in Rafols & Meyer, 2010) both in citing and cited dimensions using the minimum necessary information.

Table 1. Interpretation of changes in the frequency distribution of external citations and associated entropy values.

Change in frequency distribution of external citation	*Change in entropy values*	*Interpretation*
Δ Number of SC's involved (Δ in the diversity of sources)	ΔH	Δ Associated multi-disciplinarity
Δ Unvenness of the distribution of citations among SC'S	- ΔH	- ΔH associated multidisciplinarity

Methodology:

Once the data regarding the disciplinary affiliation of citing and cited journals in JCR Social Sciences Edition 2011, the indicator detailed in this section was applied to the journals belonging to the first quartile of JCR-SSCI 2011 Library and Information Science. If the journal on which the indicator is being applied is classified in Information Science & Library Science and it gets 17 citations from journal B which is classified in Information Science & Library Science as well as in Geography, those citations will be counted as internal citations (Boolean "or"). The following table reflects the interpretation of the changes in the entropy values for the denominator of the indicator.

Indicator formulation:

Entropy-Based Disciplinarity Indicator (EBDI):

$$EBDI = \frac{\%IC}{\%H_{MAX(EC)+1}}$$

$\%\boldsymbol{IC}$ (Percentage of internal citations) is the percentage of citations/refs from/to journals classified at least **in the same subject category** as the unit on which the indicator is being calculated (percentage of internal citations).

$\%\boldsymbol{H_{MAX(EC)+1}}$ is the percentage that the entropy associated to the distribution of **external citations** (from a different subject category) represents respect the maximum possible entropy associated to the distribution of external citations (ln n, n being the maximum number of possible SC's in the system, adapted from Leydesdorff & Rafols, 2011).

Table 2: suggested taxonomy of studied units' role according to the combination of categorized levels of EBDI.

*Degree of disciplinarity: **citing** dimension.*	*Degree of disciplinarity: **cited** dimension.*	*Suggested interpretation of the "role" of the journal in the Sc*
HIGH	HIGH	The journal might be at the disciplinary core of its SC.
LOW	HIGH	The journal can be considered a "knowledge importer".
HIGH	LOW	The journal can be considered a "knowledge exporter"
LOW	LOW	The journal research front is not clearly in that SC. Its thematic relation to the SC is rather tangential.

The interpretation exemplified in this case is only valid for journals; other units would entail different interpretations (for example, in the case of a subject category, the "degree of disciplinary isolation" could be the axis for the interpretation).

Results:

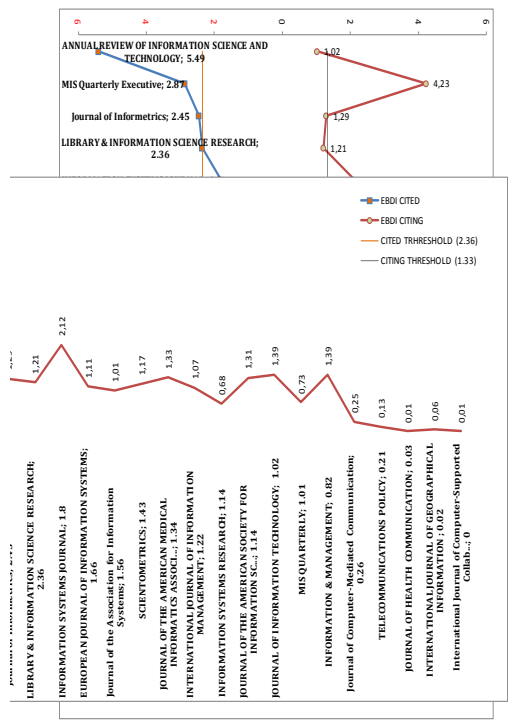


Chart 1: Cited and citing indicators associated to journals in the 1st quartile (FI) of Information Science & Library Science SC in JCR Social Sciences Edition 2011

As it can be observed in Chart 1, the Annual Review of Information Science and Technology is a clear case of knowledge input multidisciplinarity; papers whose journals are classified in a wide variety of SC's are cited by the authors publishing in this journal, but the papers published by the journal are mainly cited by authors publishing in other Information Science & Library Science journals: it might be a "knowledge importer".

MIS Quarterly Executive is highly disciplinary in both dimensions, cited and citing. Coherently with previous studies' results using other indicators (Leydesdorff and Rafols, 2011, Op. Cit.), this journal is highly monodisciplinary and might be taken as a core journal, from a knowledge classification perspective. It takes knowledge from its SC and transforms it into something mainly interesting for researchers publishing in that SC.

The Journal of Informetrics, though its values in the citing and cited dimensions are very close to the quartile threshold, could be placed in the taxonomy as a "knowledge importer", as in the case of the Annual Review of Information Science and Technology, since it is highly disciplinary in the cited dimension, but multidisciplinary in the citing dimension.

Information Systems Journal might be a clear example of "knowledge exporter", since it is disciplinary in the cited dimension but multidisciplinary in the citing dimension. The journal knowledge input frontier is mainly in its own SC, but the research published in the journal becomes interesting for researchers in other disciplines. In the same profile, it is possible to identify the Journal of Information Technology and the journal Information & Management.

References

National Academies. (2005). Facilitating interdisciplinary research. Washington, DC: National Academies. In Press.

Leydesdorff, L., & Rafols, I., (2011). Indicators of the interdisciplinarity of journals: Diversity, centrality, and citations. *Journal of Informetrics*, 5(1), 87–100

Rafols, I., & Meyer, M. (2010). Diversity and network coherence as indicators of interdisciplinarity: case studies in bionanoscience. *Scientometrics*, 82(2), 263–287.

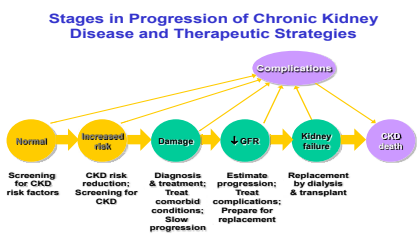
Wagner, C. S., Roessner, J. D., Bobb, K., Klein, J. T., Boyack, K. W., Keyton, J., Rafols, I., Börner, K. (2011). Approaches to understanding and measuring interdisciplinary scientific research (IDR): a review of the literature. *Journal of Informetrics*, 5(1), 14-26.

THE EPIDEMIC OF RENAL DISEASE –AN EVALUATION OF STATUS (2005-2009)

Mona Gupta[1], Divya Srivastava[2], Sandhya Diwakar[3], Arvind Singh Kushwah[4]

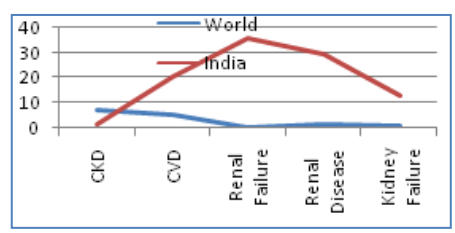
[1]*gmona7@gmail.com003*, [2]*drdivya.srivastava@gmail.com*,
[3]*sandhyadiwakar@gmail.com*, [4]*arvindsinghmca@yahoo.com*
Scientist B, Indian Council of Medical Research, New Delhi 110029 India

Renal Disease is defined as a slow lose of renal function over time. This leads to a decreased ability to remove waste products from the body and perform homeostatic functions. There are four stages in Renal Diseases which are following as:



The symptoms of renal diseases can be understood by following means:

The literature survey has indicated that there is no comprehensive work has been done by any researcher on this topic. Therefore the present study would concentrate on the work being carried out by Indian R & D scientist's vies-a-visa Global researchers. The basic data for the analysis has been culled out from MEDLINE by using search string kidney disease. There are total 2, 90,934 papers indexed worldwide out of which 5740 are Indian papers during 2005-2009. For mapping the data suitable analytical software's will be used. The main Research Areas of CKD: Cronical kidney Disease, Renal Disease, Kidney failure, Renal Failure, Renal Disease and Chronic kidney disease. In field of Cronical Kidney Disease there are total 1.20% workdone worldwide out of which India has its share of 7.53%, Cardiovascular disease research share in world context is 20.45% while out of which 5.57% papers are Indian, Renal Failure has maximum publication during 2005-09(36.13%) while India has share of .49% only. There are 29.28% research in field of renal failure in world context out of which 1.54% share is from India. There are 13.04% research is on Kidney Failure Including Indian research share of 1.02%



Publication Growth of Indian Papers in World Context during 2005-09

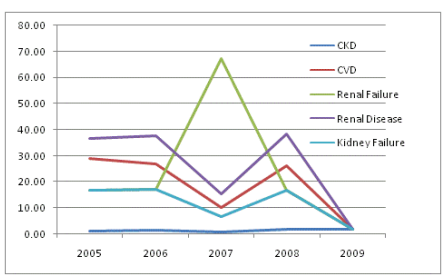
Research Progress of India in World Context:

There are 40942 papers published in 2005 followed by 42213(2006), 110650(2007), 47846(2008) and 49283 during 2009.

Research Areas	2005	2006	2007	2008	2009	Total Papers
CKD	490	601	703	800	911	**3505**
CVD	11812	11244	11320	12530	12600	**59506**
Renal Failure	6805	7214	74258	8077	8419	**104773**
Renal Disease	15030	15940	16941	18362	18934	**85207**
Kidney Failure	6805	7214	7428	8077	8419	**37943**
Total	**40942**	**42213**	**110650**	**47846**	**49283**	**290934**

Growth of Publication in different research areas during 2005-09

There is maximum research in Renal failure.



Average Growth of research areas during 2005-09

The Compound average growth shows that there is a delectations in all fields can be seen by following table.

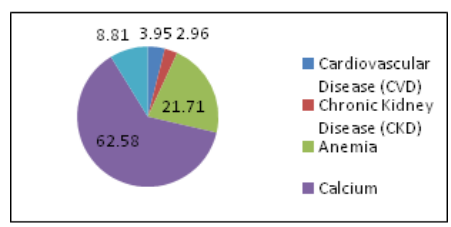
Research Areas	CAG
CKD	-96.6515
CVD	-96.2042
Renal Failure	-95.6189
Renal Disease	-94.7724
Kidney Failure	-95.6189
Total Papers Year Wise	93.9152

Research profile of productive indian authors in medicine:

The research activities of the most productive 15 Indian authors in Renal Disease published 135 and above papers during 2005-09. Of these, six authors are affiliated to AIIMS, New Delhi, two to National Institute of Immunohaematology, Mumbai, and one each to other institutions. These 10 most productive authors together contributed 575 papers in renal field during 2005-09, with an average of 5.75 papers per author and witnessed the growth of 20.86% for the papers published from 2005-09. Eight authors published higher number of papers than the group average.

Keyword Analysis:

Calcium was topmost keyword followed by Anemia, Phosphorus, Cardivascular Disease and Chronik Kidney Disease



% of Keywords used during 2005-09

Medical colleges:

Research productivity and impact of top 31 most productive Indian medical colleges during 2005-2010

Institute	TP	TC	ACPP	H-index
AIIMS, New Delhi	650	1200	1.85	39
SGPIMS, Lucknow	567	1000	1.76	35
Maulana Azad Med Coll Delhi	500	930	1.86	31
KMC, Manipal	450	805	1.79	27
BHUIMS, Varanasi	425	726	1.71	24
Shree Chitra Tirunal Inst Med Sci Tech	398	696	1.75	23
JIPMER Puducherry	312	650	2.08	20
Uni Coll Med Sci Delhi	200	595	2.98	19
Govt Med Coll Hosp Chandigarh	158	550	3.48	15
Pt. B D Sharma Post Grad Inst Med Sci	111	413	3.72	13

To conclude, CKD, CVD, Renal Failure, Renal Disease and Kidney Failure is a problem of epidemic proportions in India, and with an increasing diabetes burden, hypertension, and growing elderly population it is going to increase even further. Managing the patient population. The money invested at this time in establishing a prevention program for all these are definitely going to give results in years to come and ultimately in the long run will still be cost-effective. This money can be utilized for other healthcare programs. However, it requires a lot of data and professional lobbying with various policymakers, the MOHFW, and the Government of India.

EUROPEAN HIGHLY CITED SCIENTISTS' PRESENCE IN THE SOCIAL WEB

Amalia Mas-Bleda[1], Mike Thelwall[2], Kayvan Kousha[3] and Isidro F. Aguillo[4]

[1] *amalia.mas@cchs.csic.es*
Spanish National Research Council (CSIC), Institute of Public Goods and Policies, The Cybermetrics Lab, C/Albasanz 6-28, 28037, Madrid (Spain)

[2] *M.Thelwall@wlv.ac.uk*
University of Wolverhampton, School of Technology, Statistical Cybermetrics Research Group, Wulfruna Street, WV1 1LY, Wolverhampton (United Kingdom)

[3] *K.Kousha@wlv.ac.uk*
University of Wolverhampton, School of Technology, Statistical Cybermetrics Research Group, Wulfruna Street, WV1 1LY, Wolverhampton (United Kingdom)

[4] *isidro.aguillo@cchs.csic.es*
Spanish National Research Council (CSIC), Institute of Public Goods and Policies, The Cybermetrics Lab, C/Albasanz 6-28, 28037, Madrid (Spain)

Introduction

The web has given academics new ways to collect and disseminate the scholarly information (Chen et al. 2009; Pitzek 2002) and they have taken advantage of this in many different ways (Barjak, 2006, Mas-Bleda & Aguillo, in press; Mas-Bleda et al, in press). Scientists' web presences have also been investigated to some extent, often through personal websites (Barjak, Li & Thelwall, 2007). A survey of UK academics found these people willing to try free new social web sites but did not ask about specific services (Procter et al., 2010) and another survey confirmed that most were positive about social web initiatives (Ponte & Simon, 2011). Another study of a convenience sample survey found that academics using one type of social media site were more likely to also use another, and that younger researchers were a little more likely to use the social web than older researchers (Nicholas & Rowlands, 2011). It is unclear, however, how wide the scholarly uptake of web tools is, which ones are used, and whether they are popular amongst the senior and most influential researchers.

Research questions

The objective of this work is to identify if an influential group of researchers, highly cited scientists working at European institutions, have different types of web presences: personal websites and research group websites as well as profiles in Google Scholar, Microsoft Academic Search (MAS), Mendeley, Academia.edu and LinkedIn. The questions guiding the research are:

- What proportion of European Highly Cited (EHC) scientists have these web presences?
- Are there differences among disciplines?
- Are younger scientists more likely to have each kind of web presence?

Method

EHC scientists were identified using the online directory of highly cited researchers and the previous version of this database, both created by ISI/Thomson Reuters (www.highlycited.com/). Only 5% were women, so Microsoft Academic Search was used to increase this percentage. A total of 1,583 EHC researchers were identified. Deceased researchers were removed to give a final population of 1,517 scientists, 1,360 (90%) men and 157 (10%) women. The scientists were then grouped into five broad disciplines: engineering, physical sciences (in which maths is included), health sciences, life sciences and social sciences. These data were taken from a previous study (Mas-Bleda et al, in press) and then updated. Finally, the web presences of the scientists were manually searched for between Nov. 2012 and March 2013.

Results

Table 1 shows the proportion of living researchers with different web presences for each discipline. Overall, this group of scientists had a low web presence, except for personal websites and for Microsoft Academic Search. Perhaps the most surprising result was the very low use of Academia, especially for researchers from hard sciences, health sciences and life sciences.

Concerning disciplines, social scientists had the most personal websites (83%), but fewest research group websites (1%), suggesting a lack of cooperation among researchers in this discipline. This group also had the largest proportion of Google Scholar profiles.

Independent samples median tests were conducted to see if the average age of the scientists having a web presence differed from the average age of the scientists not having a web presence. EHC researchers with a personal website or a profile in Google Scholar or LinkedIn, tended to be younger than those who did not (Sig. 0.000 for the first two and Sig. 0.006 for LinkedIn). The difference in medians was not large, however, at only three years in most cases except for personal websites (6 years) and Google Scholar profiles (6 years).

Table 1. Percentage of EHC scientists with different web presences for each discipline.

Web presences	Discipline				
	Eng. (n=241)	Physical (n=353)	Health (n=435)	Life (n=413)	Soc. (n=75)
Personal website	78.0%	76.5%	53.6%	52.5%	82.7%
Res. group website	12.4%	14.7%	18.2%	22.0%	1.3%
G. Scholar	15.4%	8.8%	6.4%	7.0%	24.0%
MAS	99.2%	99.2%	98.4%	98.5%	97.3%
Mendeley	6.2%	4.2%	6.0%	7.7%	8.0%
Academia	4.1%	0.6%	0.7%	0.7%	5.3%
LinkedIn	27.0%	18.4%	29.0%	19.6%	25.3%

Eng. = Engineering; **Physical** = Physical sciences
Health = Health sciences; **Life** = Life sciences
Soc. = Social sciences

Chi-square tests were also conducted to for significant differences between types of web presence for each discipline. In engineering, researchers with profiles in LinkedIn also tended to have personal website and profile in Google Scholar and Academia ($p. < 0.05$), and those with profiles in Google Scholar also tended to have profiles in Mendeley and LinkedIn.

In the physical sciences, researchers with Google Scholar profiles also tended to have personal websites and profile in Mendeley. In health sciences, researchers with personal websites tended to also have Google Scholar and LinkedIn profiles. In life sciences, researchers with profiles in Mendeley

also tended to have profiles in Google Scholar and LinkedIn. In social sciences, researchers with profiles in Google Scholar also tended to have profiles in Academia and LinkedIn.

Conclusions

The results show that EHC scientists tend to have personal websites (created by themselves or by their institution) and Microsoft Academic Search profiles (created by Microsoft), but few have created a profile in any of the other academic sites investigated. Perhaps surprisingly, the only non-academic site in the list, LinkedIn, was the most popular social web site. Nevertheless, this presence is much less widespread than it found in a previous study (Bar-Ilan et al, 2012).

Unsurprisingly, however, there were disciplinary differences in the most popular types of site used and younger researchers were more likely to have a presence than older researchers in most types of site. Finally, researchers having one type of profile were more likely to have another in many cases, suggesting that scientists tend not to restrict themselves to just one type of social web presence, if they choose to create one at all.

Acknowledgments

This work is supported by ACUMEN (Academic Careers Understood through Measurement and Norms) project, grant agreement number 266632, under the Seventh Framework Program of the EU.

References

Barjak, F. (2006). The role of the Internet in informal scholarly communication. *Journal of the American Society for Information Science and Technology*, 57(10), 1350-1367.

Barjak, F., Li., X. & Thelwall, M. (2007). Which factors explain the web impact of scientists' personal homepages? *Journal of the American Society for Information Science and Technology*, 58(2), 200-211.

Bar-Ilan, J., Haustein, S., Peters, I., Priem, S., Shema, H. and Terliesner, J. Beyond Citations: Scholars' Visibility on the Social Web. In *Proceedings of 17th International Conference on Science and Technology Indicators* (pp. 98-109), Montréal: Science-Metrix and OST.

Chen, C., Sun, K., Wu, G., Tang, Q., Qin, J., Chiu, K. et al (2009). The impact of internet resources on scholarly communication: a citation analysis. *Scientometrics*, 81(2), 459-474.

Mas-Bleda, A. & Aguillo, I. (in press). Can a personal website be useful as an information source to assess individual scientists? The case of European highly cited researchers. *Scientometrics*.

Mas-Bleda, A. Thelwall, M., Kousha, K. & Aguillo, I. (in press). Successful researchers publicizing research online: an outlink analysis of European highly cited scientists' personal websites. *Journal of Documentation*.

Nicholas, D. & Rowlands, I. (2011). Social media use in the research workflow. *Information Services & Use*, 31, 61-83.

Pitzek, S. (2002). *Impact of online-availability of science literature*. Retrieved March 7, 2013 from: http://www.vmars.tuwien.ac.at/courses/ proseminar/doc/paperserver.pdf.

Ponte, D. & Simon, J. (2011). Scholarly communication 2.0: Exploring researchers' opinions on Web 2.0 for scientific knowledge creation, evaluation and dissemination. *Serials Review*, 37(3), 149-156.

Procter, R., Williams, R., Stewart, J., Poschen, M., Snee, H., Voss, A. & Asgari-Targhi, M. (2010). Adoption and use of Web 2.0 in scholarly communications. *Philosophical Transactions of The Royal Society A*, 368(1926), 4039-4056.

EVALUATING THE INVENTIVE ACTIVITY OF FOREIGN R&D CENTERS IN ISRAEL: LINKING PATSTAT TO FIRM LEVEL DATA

Daphne Getz[1], Eran Leck[2] and Amir Hefetz[3]

[1] *Daphne@sni.technion.ac.il*
Senior Research Felllow, Samuel Neaman Institute for Advanced Studies in Science and Technlogy, Technion – Israel Institute of Technoloy, Technion City, Haifa 32000 (Israel)

[2] *leck@sni.technion.ac.il*
Researcher, Samuel Neaman Institute for Advanced Studies in Science and Technlogy, Technion – Israel Institute of Technoloy, Technion City, Haifa 32000 (Israel)

[3] *ahefet01@campus.haifa.ac.il*
Researcher, Samuel Neaman Institute for Advanced Studies in Science and Technlogy, Technion – Israel Institute of Technoloy, Technion City, Haifa 32000 (Israel)

Abstract

The state of the art in patent statistics today involves linking patent data to complementary databases in an attempt to supply additional information on the patent's assignees. Notable examples to these types of databases are the plug & play extensions to PATSTAT – OECD HAN, the EEE-PPAT tables, OECD REGPAT and the OECD ORBIS database (business register data). In this research we set out to investigate the scope of inventive activity conducted by multinational companies that established local branches in Israel. The Israel Venture Capital (IVC) database was linked to PATSTAT in order to analyse the inventive activity of these firms. The outcome of this exercise resulted in the development of new globalization indicators and the attainment of high resolution data on the inventive characteristics of foreign R&D centres in Israel.

Methodology

- The unit of measurement for inventive activity is "*distinct invention"* – the earliest (priority) filing of the same application anywhere in the world.
- The distinct invention indicator is based on the DOCDB family and is aimed at neutralizing double counting of identical patent applications (inventions), as a result of their filing in different patent offices.
- An improved version of KUL's EEE-PPAT tables (assignee harmonization and sector allocation), for PATSTAT was used to identify all foreign owned applications attributed to the business sector (Du Plessis et al., 2009).
- The resulting data subset was linked to firm level data (IVC) encompassing rich information on the characteristics of 264 foreign R&D centres in Israel.

- We are interested in identifying and analysing inventions filed by multinational companies that established local branches ("foreign R&D centres") in Israel. The ownership of these inventions is foreign, while the inventors are Israeli.

Research Findings

- In the past decade the inventive activity of foreign R&D centres has risen by 144% (Figure 1).

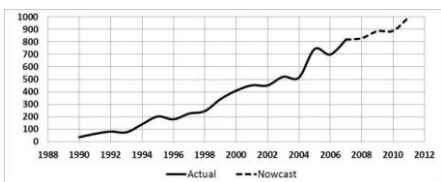


Figure 1 Number of distinct inventions filed by foreign R&D centres.

- Distinct inventions filed by IBM, SanDisk and Intel constituted 39% of the total inventive activity of foreign R&D centres in Israel during the 2006-2010 time period (Table 1).

Table 1: Distinct inventions filed by foreign R&D centres (top assignees).

R&D centre	2001-2005	2006-2010
IBM	491	463
SanDisk	75	394
Intel	484	321
HP	172	168
Microsoft	66	142
Qualcomm	55	121
All R&D centres	2679	3016

- In 2011, the inventive activity of foreign R&D centres in Israel constituted 27% of total Israeli distinct inventions (Figure 2) and 61% of total foreign owned distinct inventions attributed to the business sector (Figure 3).

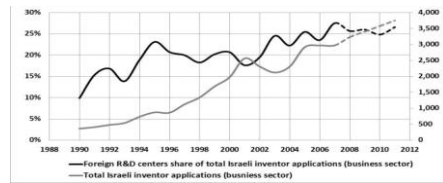


Figure 2: Foreign R&D centres' share of total Israeli distinct inventions (business sector).

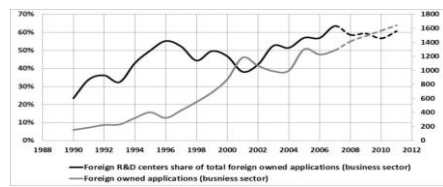


Figure 3: Foreign R&D centres' share of total foreign owned applications (business sector).

- 75% of the inventive activity of foreign R&D centres (Figure 4) is conducted by well established firms (more than 30 years, e.g. Intel, IBM) or by new R&D centres in Israel (1-10 years, e.g. Qualcomm, Samsung).

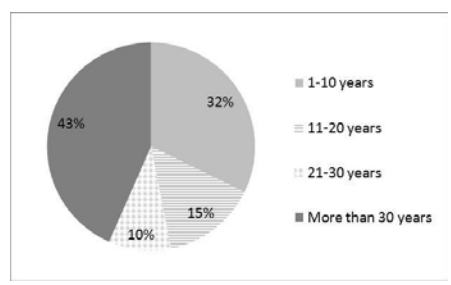


Figure 4.Distribution of distinct inventions filed by foreign R&D centres by years of activity in Israel.

- More than 95% of distinct inventions are attributed to the high technology and medium-high technology sectors (Figure 5).

- In the 1990-2010 time period, at least 1361 distinct inventions were transferred from the ownership of Israeli companies or start-ups to the possession of foreign R&D centres (MNCs), due to acquisitions or mergers (Figure 6, Table 2).
- These 1361 distinct inventions constitute approximately 13.5% out of the total patent portfolio of the R&D centres.

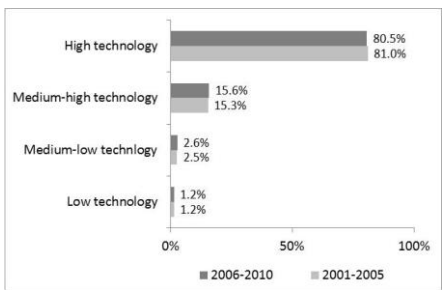


Figure 5: Distribution of distinct inventions by technological intensity.

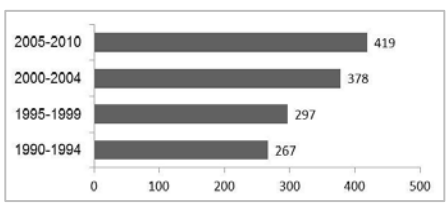


Figure 6: Number of distinct inventions acquired from Israeli firms by foreign R&D centres as a result of acquisitions or mergers.

Table 2: Distinct inventions acquired from Israeli firms (top acquirers).

Name prior to acquisition	Current R&D centre affiliation	Distinct inventions
Indigo	HP	134
Medingo	Roche	70
Aladdin	Safenet	60
M-Systems	SanDisk	53
Anobit Tech.	HDC Apple	45

Conclusions

In the past decade, the rate of transfer of Israeli IP, know-how and technology to the possession of foreign R&D centres has substantially increased.

- There is a significant rise in the absolute number of distinct inventions filed by foreign R&D centres and in their respective share out of total Israeli inventive activity.
- Increasing trend of obtaining Israeli IP by means of acquisition of Israeli firms and start-ups. Acquired patents are becoming a substantial share out of the total patent portfolio of foreign R&D centres in Israel.
- These trends should be taken into account in the evaluation of Israel's national R&D policy.

References

Du Plessis, M., Van Looy, B., Song, X., & Magerman, T. (2009). Data production methods for harmonized patent indicators: Assignee sector allocation. *EUROSTAT Working Paper and Studies*, Luxembourg.

EVALUATION OF RESEARCH IN SPAIN: BIBLIOMETRIC INDICATORS USED BY MAJOR SPANISH RESEARCH ASSESSMENT AGENCIES

Alicia F. Gómez-Sánchez[1] and Rebeca Isabel-Gómez[2]

[2] *afgomez@cnic.es*
Fundación Centro Nacional de Investigaciones Cardiovasculares (CNIC).
C/ Melchor Fernández Almagro, 3 28029 Madrid (Spain)

[2] *rebeca.isabel.ext@juntadeandalucia.es*
Agencia de Evaluación de Tecnologías Sanitarias de Andalucía (AETSA).
Avda. Innovación. Edificio ARENA . 41020 Sevilla (Spain)

Introduction

One of the most important current applications of bibliometrics is assessment of research, and bibliometric indicators can be considered as tools for the evaluation of the scientific productivity of an individual researcher, a group or an institution.

Taking this as a starting point, we would like to check who is setting the patterns for scientific output evaluation in Spain in the area of biomedicine: What indicators have been used in recent years by funding and evaluating institutions in Spain? Are these institutions appropriately exploiting the resources provided by the bibliometry? What factors should be taken into account when defining indicators? What are the most accurate indicators for measuring research and performance? In brief, our objective is to observe the indicators and criteria that are being used by the main Spanish Agencies for the evaluation of researchers and institutions in Spain.

Methods

This study analyses the evaluation criteria and indicators used by the major agencies of the Spanish research evaluation system. They are the National Assessment and Planning Agency (ANEP), the National Evaluation Commission for Research Activity (CNEAI) and the National Agency for Quality Assessment and Accreditation (ANECA). We compared the indicators used with the main characteristics of scientific communication and publications in the area of health sciences and we make some recommendations for measuring science in a more accurate way.

Research Evaluation in Spain: Results

Research assessment in Spain for a long time had two broad objectives: the pre-evaluation of research projects for financing and the external evaluation of the research activity of individual researchers over six-year periods, called *sexenios*. In recent years, we have seen the introduction of other kinds of evaluation, for example to provide accreditation to institutions of excellence or simply to measure research activity.

Regarding the indicators and criteria used, the quantity of publications is the most demanded criteria for being evaluated. The majority of the evaluation programs in Spain consider

absolute data (number of publications, IF, citations, etc.), do not take into account normalized indicators.

Table 1. Main objectives of Spanish Agencies for evaluation of researches.

ANECA Since 2002	CNEAI Since 1989	ANEP Since 1986
• Offers external quality assurance to the university system. • Evaluates and accredits university lecturers in order to integrate the Spanish system into the European Higher Education Area	• Performs an annual evaluation of research activity of university researchers and scientists in the CSIC (Spanish National Research Council) • Incentivizes research work and its diffusion both nationally and internationally	• Evaluates scientific & technical quality of proposals seeking public funding or financing • Improves the capacity of the public Science and Technology system

Types and proportions or collaborations, as well as productivity calculated by counting the number of publications per person and year. The impact factor (IF) of the journals is other of the most important criteria used. The most common database used in Spain is the ISI WoS.

Table 2. Main Indicators used by Spanish Agencies for evaluation of researches.

Indicators used	ANECA	CNEAI	ANEP
Output	✓	✓	✓
Normalizad Impact			✓
High Quality Publications (Q1/D1)		✓	✓
Leadership			✓
Cites per Document		✓	✓

Conclusions and recommendations

The employment of excellence indicators (10% most cited papers in their respective fields or in top-journals), should be more extended. Moreover, other aspects as leadership and visibility should be also bear in mind.

For measuring individual investigators other indicators as the h-index or normalized indicators as the crown or the SNIP should be recommendable. Moreover, self-citations are usually not considered and much less uncitedness, which is not considered at all. It would be also interesting to think about the number and percentage of documents cited and not cited.

In order to have a more global and complete evaluation it would be essential to combine different indicators and develop an evaluation program allowing a multidimensional, comprehensive assessment, depending on the needs and objectives of the assessment. Furthermore, differences should be taken into account within different subject areas. For instance, as indicated by the ANEP, in the context of health sciences it should be considered if the results of basic research and preclinical are transferred to clinical or applied science (e.g. through clinical guidelines) and to innovation. To that effect, the number of patents or utility models should also be taken into consideration.

Recent recommendations, as the "San Francisco Declaration on Research Assessment", confirm the high importance to review the way scientific research is evaluated.

As a final point, other alternative sources and toolkits, for instance the F1000 Journal Rankings, ALTmetrics or Article-Level Metrics, should be also taken into account as an alternative approach to evaluate the scientific impact of scholarly communications.

Finally, the way scientific production is measured in Spain seam not to be really accurate and the instruments used are not taking advantage of the real evolution of Bibliometrics science. Perhaps current development of bibliometric units in universities and research institutions could help to

maximize benefits from the bibliometrics advances.

References

ANECA. (2013).*Agencia Nacional de Evaluación de la Calidad y Acreditación (ANECA).* Retrieved january, 2013 from http://www.aneca.es/ANECA

ANECA. (2008). *Programa ACADEMIA. Principios y Orientaciones para la Aplicación de los. Criterios de Evaluación*. Retrieved january, 2013 from http://www.aneca.es/content/download/10527/118089/version/1/file/academia_14_ppiosyorientaciones.pdf

ANECA. (2010). *Informe sobre el estado de la evaluación externa de la calidad en las universidades españolas 2010.* Retrieved january, 2013 from www.aneca.es/content/download/12369/152900/file/informe_calidadenunis10_120312.pdf

ANECA. (2011). *Programa de evaluación de profesorado para la contratación: Guía de ayuda al solicitante*. Retrieved january 2013 from http://www.aneca.es/content/download/12045/135410/file/pep_guiadeayuda_120118.pdf

Ministerio de Educación Cultura y Deporte. (2013). *Comisión Nacional Evaluadora de la Actividad Investigadora (CNEAI)*. Retrieved january 2013 from http://www.mecd.gob.es/ministerio-mecd/organizacion/organismos/cneai.html

Huggett, S. (2012). F1000 Journal Rankings: an alternative way to evaluate the scientific impact of scholarly communications. *Research Trend, 26*, 7-10.

Ministerio de Educación Cultura y Deporte. (2012). *Resolución de 19 de noviembre de 2012, de la Comisión Nacional Evaluadora de la Actividad Investigadora, por la que se establecen los criterios específicos en cada uno de los campos de evaluación*. Retrieved january 2013 from http://www.boe.es/boe/dias/2012/11/29/pdfs/BOE-A-2012-14633.pdf

Ministry of Economy and Competitiveness. (2012). *National Evaluation and Foresight Agency (ANEP)*. Retrieved january 2013 from http://www.idi.mineco.gob.es/portal/site/MICINN/menuitem.29451c2ac1391f1febebed1001432ea0/?vgnextoid=3cb39bc1fccf4210VgnVCM1000001d04140aRCRD

San Francisco Declaration on Research Assessment. Putting science into the assessment. Recommendations of editors and publishers of scholarly journals during the Annual Meeting of The American Society for Cell Biology (ASCB) in San Francisco, CA, on December 16, 2012.

Sanz-Menéndez, L. (1995). Research actors and the state: research evaluation and evaluation of science and technology policies in Spain. *Research Evaluation, 5*(1), 79-88.

AN EXPERIENCE OF THE INCLUSION A NEW METHODOLOGY IN SELECTING THE REVIEWERS FOR GRANT APPLICATIONS

Primoz Juznic[1], Robert Matoh[2] and Doris Dekleva Smrekar[3]

[1] *primoz.juznic@ff.uni-lj.si*
University of Ljubljana, Faculty of Arts, Department of Library, Information Science and Book Studies, Aškerčeva 2, 1000 Ljubljana (Slovenia)

[2] *matoh1@siol.net*
Biomedical Research and Innovative Society, *Puhova* ulica 10, 1000 Ljubljana (Slovenia)

[3] *doris.dekleva@ctk.uni-lj.si*
Central Technological Library at the University of Ljubljana, Trg republike 3, 1000 Ljubljana (Slovenia)

Introduction

In Slovenia, to each annual Call for research projects proposals by the Slovenian Research agency about 1,000 applicants apply. The project selection process takes place in two stages. In the evaluation of projects in first phase several criteria are used (Juznic et al, 2010), a combination of bibliometric indicators and peer reviews. In the second phase only peer review evaluation in the form of rating, is used. It is important to provide an effective and transparent process of selecting reviewers, while it is the most part of the evaluation of research projects (Lee et al, 2013). Modern ways of seeking reviewers should provide intellectual awareness of the reviewers, the exclusion of the conflict of interests or professional association and it is desirable a good responsiveness of the invited reviewers. (Harley and Acord, 2011)

To evaluate the peer review procedure we analysed the relationship between bibliometric methods and expert evaluation of proposed projects, the response of the reviewers and as presented in this paper, a degree of reliability of the referees' grades using the "intraclass correlation" method.

Materials and Methods

Applicants have been divided by the six major research disciplines within the sciences according to Slovenian Research Agency Field of research Classification (Natural sciences and mathematics, Engineering sciences, Medical sciences, Biotechnical sciences, Social sciences and Humanities), further on into greater number of sub-fields and finally by thematically related topics into "clusters" of up to five applications. Each project were suppose to have three reviewers.

Main tool for searching and the selection of reviewers was Reviewer Finder**.** Basically Reviewer Finder uses a semantic search, but it also provides a list of potential reviewers with indicators of expertise and possible conflicts of interest by entering the details from the application of entire cluster (http://info.scival.com/reviewer-finder). The list of potential reviewers

were filtered by the criteria of the researchers publishing (h-index, their review publications, publishing in recent years), as well as geographic origin (Europe) to maximize their responsiveness.

Using Reviewer Finder 544 reviewers were selected. The goal was to find same three reviewers for the evaluation of one "cluster" of similar projects. Although it was not possible for all reviewers who were eligible for the assessment for more than two projects we found.

The biggest share of reviewers, who accepted the Agency´s invitation, came from Portugal, Finland and Greece (over 40%). The Italian reviewers were third most numerous in regard of overall sent invitations for participating in this Peer evaluation (47), only behind UK (89) and Germany (76), but their responsiveness score was also above average with 39%. The poorest response was from some Scandinavian countries (Sweden and Denmark) and France and Switzerland.

One possible assumption, which has arisen in this case was, that the researchers with higher h-index would more likely to refuse the participation in peer evaluation. T-test revealed that in the majority of fields h-index of a researcher was not a determining factor in the decision to enter peer review process.

In the research we wanted to assess the reliability of the reviewers grades. To meet this aim, a measure of Intraclass Correlation Coefficient (ICC) was used (Nichols, 1998). This is a tool, a general measurement of agreement or consensus, where the Coefficient represents agreements between two or more raters or evaluation methods on the same set of subjects. ICC has advantages over correlation coefficient, in that it is adjusted for the effects of the scale of measurements, and that it will represent agreements from more than two reviewers. The goal was to compare the reliability of grades within the same set of reviewers.

A total of 313 project proposals from 61 scientific disciplines were included in the study, where each discipline was considered its own unit of the analysis. In other words, the main goal was to measure absolute agreement between raters/reviewers on each scientific field. Because the same sets of reviewers were not always used in peer evaluation inside one discipline (especially when there were more than four or five applicants), we must also stress, that according model of ICC was used. In that way, two coefficients were calculated: beside above mentioned coefficient of absolute agreement, which is an index for the reliability of different reviewers averaged together, the index for the reliability of the ratings for one, single reviewer, usually lower of the two, computed by MedCalc statistical software, was calculated.

Results and Discussion

Positive correlation coefficients were discovered in 46 out of 59 disciplines, which is almost 15% better result than in the previous call of proposal, where Reviewer finder was not used as extensively**.**

Concerning the level of the reliability, it has been revealed, that the border of ICC=0,300, which in our case is considered as a solid level of reliability/agreement among reviewers, was surpassed by 76,09% of the disciplines, compared to 71,42% from previous period.

The reviewers searching process using Reviewer Finder, generally turns adequate for searching same reviewers for more than one project. Their response was satisfactory and the analysis also showed a high degree of

reliability assessment for the majority of fields.

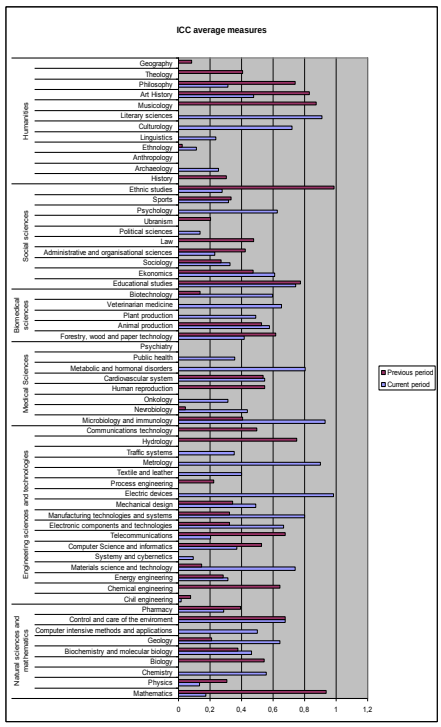


Figure 1. The ICC average measures by research fields for the current and previous period

References

Harley, D. and Acord, S. K. (2011). Peer Review in Academic Promotion and Publishing: Its Meaning, Locus, and Future. Retrieved September 31, 2012 from: http://escholarship.org/uc/item/1xv148c8#page-1

Lee, C. J., Sugimoto, C. R., Zhang, G. and Cronin, B. (2013). Bias in Peer Review. *Journal of the american society for information science and technology*, 64(1), 2–17.

Nichols, D. P. (1998). *Choosing an intraclass correlation coefficient*. Retrieved December 13, 2011 from: http://www.sph.emory.edu/observeragreement/spss.pdf .

Juznic, P. , Peclin, S. , Zaucer, M., Mandelj, T. , Pusnik, M. and Demsar, F. (2010). Scientometric indicators: peer-review, bibliometric methods and conflict of interests. *Scientometrics*, 2, 429-441.

EXPLORING INTERDISCIPLINARITY IN ECONOMICS THROUGH ACADEMIC GENEALOGY: AN EXPLORATORY STUDY

Chaoqun Ni[1] and Cassidy R. Sugimoto[1]

[1] *{chni, sugimoto}@indiana.edu*
School of Library and Information Science, Indiana University Bloomington, Bloomington, IN 47405 (USA)

Introduction

Interdisciplinarity has been heavily studied by scientometricians (e.g., Leydesdorff, 2007), mostly relying on ISI-indexed journals and intercitations between journals and JCR subject categories as indicators of interdisciplinarity (e.g., Cronin & Meho, 2007). Few studies have examined "author-based indicators of disciplinarity" (Sugimoto, Ni, Russell, & Bychowski, 2011) and those that do, typically focus on co-authorship and disciplinary affiliation of the author to classify disciplinarity (e.g., Schummer, 2004). However, there are several limitations to such an approach: journal articles are not the sole genre of scholarly dissemination and are likely to skew impressions of disciplinary interaction due to the wide difference in production rates; and using collaboration neglects the many disciplines for which sole authorship is still the norm (Ni, Sugimoto, & Jiang, 2013).

Doctoral dissertations provide a useful alternative for scientometric research. All research disciplines produce dissertations; therefore, this genre does not favour certain disciplines. Each individual produces only a single dissertation in each discipline; therefore, dissertations are not skewed in the direction of subdomains or authors who might be inordinately prolific. Finally, dissertations provide the opportunity to study mentoring through advisorship, a relatively unexplored network structure for scientometric research (Russell & Sugimoto, 2009).

This paper extends previous work done in the areas of LIS and Sociology (Sugimoto, Ni, Russell, & Bychowski, 2011; Ni & Sugimoto, 2012, 2013) by using the approach of academic genealogy to explore the interdisciplinary composition of Economics (Sugimoto, Ni, Russell, & Bychowski, 2011; Ni & Sugimoto, 2012; 2013).

Data Collection & Processing

This paper relies on the dissertation data provided by ProQuest database (hereafter PQuest). PQuest covers over 2.3 million dissertations from about 1,490 institutions across 66 countries from the last two centuries. This project utilized a subset of the PQuest: those listed in the Economics Subject Category. Thus, any dissertations in PQuest with one of their SCs belonging to Economics were considered as an Economics dissertation—thus allowing the SC to serve as a proxy for a discipline. As shown in figure 1, each row is a PQuest Subject Category (hereafter SC), and a dissertation can be assigned to multiple SCs. A dissertations can be assigned with single

or multiple SCs, and multiple SCs might belong to the same discipline.

Results Analysis & Conclusions

Overview of Economics in PQuest

Economics is a discipline with long history and large number of graduates. In PQuest, the earliest Ph.D. dissertation was finished at the year of 1899 at Yale University. There are 76,336 dissertations of Economics in PQuest, distributed unevenly across decades, as shown in table 1. Additionally, one of the most important information used in this project, the advisor of each dissertation, is not available for all dissertations. As shown in table 1, there are in total 41,317 (54.13%) dissertations which provide advisor information, and majority of which concentrate on the most recent two decades.

Table 1. Number of Economics Dissertation &Advisor-available Dissertation by Decade

Decade	#Diss	#Diss_Advisor	%
≤1900'S	2	0	0.00%
1910'S	2	1	50.00%
1920'S	4	0	0.00%
1930'S	30	0	0.00%
1940'S	93	0	0.00%
1950'S	1976	11	0.56%
1960'S	6903	0	0.00%
1970'S	10809	4	0.04%
1980'S	12923	2599	20.11%
1990'S	20687	17364	83.94%
≥2000's	22907	21338	93.15%
Total	76336	41317	54.13%

The 76,336 dissertations were finished by graduates of 497 institutions across 22 countries, of which 98.21% were finished in institutions located in North America. This is probably due to the fact that ProQuest itself is North American dominant.

Overview of Economics Advisors in PQuest

As mentioned above, only 41,317 dissertations provided advisor information in PQuest, i.e. 15,191 unique advisors. Only 4,544 (23.67%) advisors were able to be identified for the dissertations (for which these advisors were granted their degrees) in PQuest. These advisors, however, account for 47.76% of the total advisorships. For those unique advisors identified, they obtained their degree from 310 institutions located in 10 countries. About 95% of these institutions are located at U.S.. Harvard University ranks first by exporting 211 advisors to the discipline of Economics, and Stanford University ranks second.

The Disciplines of Economics Advisors

Economics advisors got degrees from 76 different disciplines (including Economics). About 82.35% have their dissertations only categorized into a single discipline, and 17.65% of advisors multiple disciplines. It seems that Economics is the major discipline exporting advisors to its own field, and Business Administration and Political Science are the second and third. Figure 1 displays the percentage of advisors who received their degrees in each decade in each of top 10 disciplines. Economics, Business Administration, Engineering, Statistics and Health Sciences are increasingly interacting with Economics over time, by exporting advisors to Economics. On the other hand, Political Science, History, and Psychology are these three disciplines exporting fewer advisors to Economics over time.

The dissertation of Economics advisors can have only Economics as its discipline (EcoOnly), Economics and other disciplines as its disciplines

(EcoMix), or non-Economics disciplines as its disciplines (NonEco). As shown in Figure 2, the percentage of advisors receiving their degrees purely in Economics has decreased in the last two decades, while that with a mixture of Economics and other disciplines are increasing. Meanwhile, the percentage of Economics advisors receiving degrees from NonEco disciplines is increasing in the last two decades.

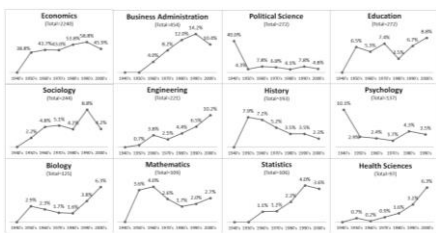


Figure 1 Percentage of Economics Advisors Exported by each Discipline by Decade

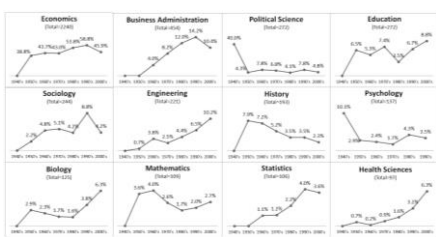


Figure 2. Percentage of EcoOnly, EcoMix & NonEco Advisors by Decade

Future Work

This is an in-progress project and only shows a preliminary result. The future plan on this project is to solve some limitations we encountered. First of all, not all dissertations have advisor information. For better diachronic studies, manual data collection of these advisors will be necessary. Second, author-name disambiguation is still not perfectly accurate and requires additional refinement. Lastly, using SCs as a proxy for disciplinarity introduces some limits to interpretations. Hence, future efforts will be made to refine these methods and add additional disciplines in order to generate a better understanding of the interaction of disciplines through academic genealogy.

Acknowledgments

This work was supported by the Faculty Research Support Program from Indiana University and the National Science Foundation (SciSIP program) (Grant No. SMA-1158670). The authors would also like to thank ProQuest for making this data available for research.

Reference

Cronin, B., & Meho, L.I. (2007). The shifting balance of intellectual trade in Information Studies. *Journal of the American Society for Information Science and Technology*, *59*(4), 551-564.

Leydesdorff, L. (2007). Betweenness centrality as an indicator of the interdisciplinarity of scientific journals. *Journal of the American Society for Information Science and Technology,* 58(9), 1303-1319.

Ni, C., & Sugimoto, C. R. (2012). Using doctoral dissertations for a new understanding of disciplinarity and interdisciplinarity.*ASIST,* Baltimore, MD.

Ni, C., & Sugimoto, C. R. (2013). Academic genealogy as an indicator of interdisciplinarity: a preliminary examination of sociology doctoral dissertations. *iConference* 2013.

Ni, C., Sugimoto, C. R., & Jiang, J. (2013). Venue-Author-Coupling: A Novel Measure of Identifying Disciplines through Author Communities. *Journal of the American Society for Information Science and Technology*. 64(2), 265–279.

Schummer,J.(2004).Multidisciplinarity,interdisciplinarityand patterns of research collaboration in nanoscience and nanotechnology. *Scientometrics*, *59*(3), 425-465.

Sugimoto, C. R., Ni, C., Russell, T. G., & Bychowski, B. (2011). Academic genealogy as an indicator of interdisciplinarity: An examination of dissertation networks in Library and Information Science. *Journal of the American Society for Information Science and Technology*. 62(9),1808-1828.

FEATURES OF INDEX TERMS AND NATURAL LANGUAGE WORDS FROM THE PERSPECTIVE OF EXTRACTED TOPICS

Ritsuko Nakajima[1] and Nobuyuki Midorikawa[2]

[1] *nakajima@u.tsukuba.ac.jp*
Graduate School of Library, Information and Media Studies,
University of Tsukuba, 1-2 Kasuga, 305-8550 Tsukuba (Japan)

[2] *midorika@slis.tsukuba.ac.jp*
Faculty of Library, Information and Media Science,
University of Tsukuba, 1-2 Kasuga, 305-8550 Tsukuba (Japan)

Introduction

Important databases are equipped with controlled indexes, which are essentially controlled vocabularies based on a thesaurus. Controlled indexing is a highly reputed system as it has many advantages such as enabling high recall and precision in search. However, it also has some drawbacks; for instance, it requires experts, which increases the cost, and it hinders quick reporting. Recently, many database services have introduced automatic indexing to solve these problems. According to Abdou & Savoy (2008), only a few studies have directly compared the performances of manual and automatic indexing methods. In this study, we extracted the information on research topics from natural language words and manually indexed terms and compared the features of each from the viewpoint of topic extraction.

Method

To understand the difference between the features of natural language and controlled vocabulary, we extracted information from a bibliographic database to perceive that research topics emerge, develop, and decline. For data sources, we used abstracts and index terms. Among previous studies on the analyses of research trends obtained from bibliographic databases, recent works based on index terms use clustering methods to gain comprehensive trends in each research field (Tseng et al., 2008; Ohniwa et al., 2010). In this study, we focused on understanding features of manually indexed terms through comparison with natural language words. We tried to extract more specific information within a relatively limited research area, which we chose as "high temperature superconductor" (HTS) and "phonon"; this is because it is clear when research in these areas began and because the literature is rich in reviews and documentation on its history, which is helpful to understand the research topics from the extracted words.

Experiment

Data

We used the CAplus data of 2,259 articles by searching the keywords "high temperature superconductor" and "phonon." Figure 1 shows the number of publications by year.

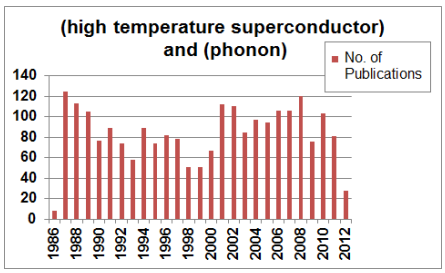


Figure 1. Number of publications containing "high temperature superconductor" and "phonon."

The number of publications repetitively increased and decreased several times, and this variation can be presumably linked with the changes in popularity of certain research fields. For the research topic extraction from abstracts as natural language and for index terms as controlled vocabulary, we considered the period from 1995 to 2012, focusing on the increase in publications after one decrease around 1998.

In CAplus, subjects are indexed in the Index Term and Supplementary Term fields. Controlled (uncontrolled) terms are contained in Index Term as "controlled vocabulary (uncontrolled vocabulary)."

Topic extraction from abstracts (natural language)

To clarify the features of topics, we used groups of two words (bigrams) as units for analysis. The bigrams were ranked by frequency of articles. Next, the following processing was performed to extract new topics that were remarkable during the period.

(a) During 1995-2000, a word contained in a bigram, which appeared more than twice a year was defined as a "base word." It was regarded as a conventional word generally used in the research field and which did not express a certain research topic.

For bigrams appearing in and after 2001, if both words in the bigrams are base words, the bigram was deleted. Bigrams containing one or two new words were ranked by frequency of articles. (Figure 2)

(b) A different method was used for choosing the base words in (a). For bigrams appearing in and after 2001, any new words in the bigram were added to the base words every year. Then, bigrams were ranked by frequency of articles.

(c) Step (a) was repeated, with the addition of stemming.

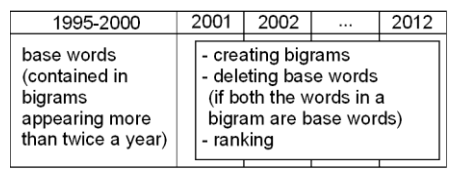


Figure 2. Outline of processing(a).

Topic extraction from Index Terms (controlled vocabulary)

Next, controlled vocabulary was employed in the field of Index Term. Both a word and a compound word were used as they were. As in the case of abstracts, index terms that appeared more than twice for the first time in and after 2001 were ranked by frequency.

Results

From the abstracts, terms related to the following topics were extracted.

(T1) Superconductor MgB_2
- first appearance in 2001 data
- (background) A metallic superconductor was discovered in 2001.

(T2) Angle resolved photoemission spectroscopy (ARPES)
- first appearance in 2001 data
- (background) The role of phonons in superconductivity came to the fore by experiments using ARPES in 2001.

(T3) Iron-based superconductor
- first appearance in 2008 data
- (background) A superconductor was discovered in 2008.

Figure 3 shows an example of the extracted terms. The number in parenthesis denotes the article frequency. The step (b) was effective in detecting new topics, whereas new topics in 2001 and subsequent developments were found by step (a). No significant difference was found by step (c).

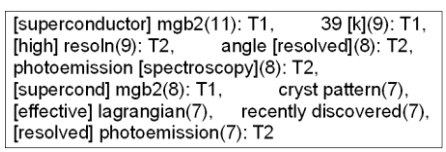


Figure 3. Top 10 terms between 2001 and 2002.

From index terms (controlled vocabulary), terms related to the following topics were extracted.
(T2) ARPES
first appearance in 2004 data
However, there are records where appropriate words are indexed as uncontrolled terms.

To confirm the association between the terms and the background, information from other resources was used. The number of citations to articles on MgB_2 and Fe-based superconductors is the highest each year, and there are many reviews on phonon measurement using ARPES (e.g. Kordyuk et al., 2010).

Conclusions

Owing to the characteristics of controlled vocabulary, information on new materials is not detectable from index terms. Therefore, we focus on the topic of ARPES. From natural language words, the emergence of the topic was detected in time when other evidence was taken into consideration. The trend that this technology was subsequently implemented frequently in this field was also observed.
As regards controlled terms, words related to ARPES appeared in 2004 for the first time. This was later than 2001, when the study on phonons in HTS, measured by ARPES, was reported and the related works increased. This is considered to correspond to the characteristics that controlled terms cannot express new topics.

References

Abdou, S., & Savoy, J. (2008). Searching in MEDLINE: Query expansion and manual indexing evaluation. Information Processing & Management, 44(2), 781-789.

Kordyuk, A. A.; et al. (2010). An ARPES view on the high-Tc problem: Phonons vs. spin-fluctuations (Review Article) The European Physical Journal Special Topics 188: 153-162

Ohniwa, R., Hibino, A., & Takeyasu, K. (2010). Trends in research foci in life science fields over the last 30 years monitored by emerging topics. Scientometrics, 85, 1–17.

Tseng, Y., Lin, Y., Lee, Y., Hung, W., & Lee, C. (2009). A comparison of methods for detecting hot topics. Scientometrics, 81(1), 73–90.

FROM CATEGORICAL TO RELATIONAL DIVERSITY – EXPLORING NEW APPROACHES TO MEASURING SCIENTIFIC DIVERSITY

Marco Schmitt[1]

[1]*mschmit4@gwdg.de*
Göttingen Centre for Digital Humanities at Georg-August-University Göttingen,
Papendiek 16, D-37073 Göttingen (Germany)

Introduction

The measurement of diversity in the sciences is a research area of crucial importance and great difficulty at the same time. The reason for this is that research diversity is a multifaceted and multidimensional concept. We will try to differentiate some of these dimensions and their interconnectedness. Special emphasis is put on concepts of relational diversity, which has a higher measurability than categorical diversity and is a good indicator of publication and citation styles in the sciences. We will discuss the different forms of relational diversity in their relation to specific scientific production fields.

The Problem of Diversity

The problem of diversity for Scientometrics and the Sociology of Science has three important dimensions. One dimension is concerned with assessing the relevance of diversity for the scientific enterprise. The second is concerned with the plasticity of the concept and the different forms it entails. And the third is concerned with the difficulty of measuring scientific diversity.

Balance of Originality and Relevance

Going back to the institutionalist analysis of science from Merton (1973) there is the theme of a tradeoff between Originality and Relevance in science. Successful science has to achieve a balance where one of the two does not drain out the other.. Diversity plays a big role in this balance, because too much diversity leads to a fragmentation with a lot of originality but less and less relevance of publications, whereas too little diversity does not provide enough originality to move forward and leads to a relevance lock-in. Therefore researchers in the Sociology of Science and Scientometrics should care about the implications of technologies or modes of governance on scientific diversity.

Different Forms of Diversity

A central problem is to identify the main features of scientific diversity because it is a concept with a lot of dimensions and an enormous plasticity. There is the diversity of access to scientific information or scientific positions, broadly speaking the theme of Diversity Studies. Then we have the epistemic diversity of scientific concepts, methods or "citational fingerprints". And we have – and we will follow this thoroughly in our approach – the diversity of the publication networks

themselves. Lastly we also have a technical diversity of instruments used to gather relevant scientific information. All these forms of diversity may vary according to personal features of the scientists, community-related features of the epistemic cultures, and overall societal developments. The question then is how to evaluate these different forms of diversity.

Problems of Measuring Diversity

The plasticity of the concept leads to severe problems of measuring scientific diversity (Havemann et al. 2007 and Heinz et al 2009). Debates on measuring biodiversity centred on three dimensions: variety, balance and disparity (Stirling 2007). In the case of the epistemic diversity of a scientific field, it is really difficult to discern the unit of measurement. The knowledge-base of such a field is not only entailed in the publications, but also in machinery and work processes. So publications can only be an indicator of epistemic diversity. But even the diversity of publications is not easy to evaluate, because they combine a lot of concepts and make a lot of different references to other publications. Just to compute the disparity between different publications seems to be a hard task.
From all these reasons the Sociology of Science and Scientometrics have to explore new ways to address the problem of diversity in science.

From Categorical to Relational Diversity

Here, we will explore what a shift from measuring categorical diversity to measuring relational diversity can achieve. While categorical diversity rests on the assumption that we can distinguish units on the basis of there inherent differences, relational diversity refers to the notion that the similarity or dissimilarity of these units rests on their structural relationships with other units (White et al 1976). Developing relational diversity based on that idea, we can go straight to measuring diversity in citation networks. In the following we will outline two possible approaches to doing that:

Network centralization

Measuring the centrality of nodes is one of the most common approaches to analysing networks (Borgatti & Everett 2005). There are a variety of centrality measures: degree, closeness, betweeness and more. Each of these single-node methods can be applied to measure the overall centralization of the whole network. All are related to the overall cohesiveness of the network and all can be easily measured. The question then is on the relation of network centralization and relational diversity? One can argue that a strong centralization and therefore a harsh core-periphery structure will inhibit diversity and should result in a low relational diversity. But cohesiveness is not always detrimental to diversity, because the flow of information between the different units is important. Strong single core networks entail a low relational diversity, they are very cohesive but peripheral information is omitted often. A less pronounced core produces a bit more relational diversity and multiple cores coupled with strong betweeness centralization are an indicator of high relational diversity in the field.

Equivalence and blockmodels

Another way to look at the problem is by applying the tools of network analysis to reduce the redundant information from networks. Structural equivalence says that every node that has the same relations to other nodes is informationally redundant and can be

viewed as having an exactly similar position (White et al 1976). We can even generalize this finding and say that a node that shares the same structural relations with another node is in the same position and reduce the network to the different positions it entails. The relational diversity of a network will then resemble the different positions for publications it entails. Their share of the whole network can account for the balance dimension in diversity and we can even calculate disparity based on the relations that differ. The reduction of the network complexity helps to identify ways to see a citation network as more or less relationally diverse.

How Relational Diversity captures Research Styles?

We will show how different research fields' exhibit vastly different relational diversity in the senses developed here. The cores and positions in their citation networks are a sign for different scientific publications styles. It is our contention that such profiles of relational diversity captures the publication-related styles of scientific fields.

Style is a concept developed by Harrison White (White 2008) that combines insights from the network research on structural equivalence and sensitivity for qualitative data on selection criteria and the production of social identities. For scientific production fields, this means that we have a network profile of cores and positions and criteria from the scientists in a scientific field to interpret what that means for their diversity.

We will show a variety of fields and their profiles and some preliminary results on the relevant criteria.

References

Borgatti, S. P., & Everett, M. G. (2006). A graph-theoretic perspective on centrality. Social networks, 28(4), 466-484.

Havemann, F., M. Heinz, M. Schmidt & J. Gläser (2007): Measuring diversity of research in bibliographic-coupling networks. In: D. Torres-Salinas und H. Moed (Hrsg*.), Proceedings of ISSI 2007,* 2, Madrid, 860–861.

Heinz, M., O. Mitesser, J. Gläser & F. Havemann (2009), Ist die Vielfalt der Forschung in Gefahr? Methodische Ansätze für die bibliometrische Messung thematischer Diversität von Fachbibliographien. In: Werner Ebeling & Heinrich Parthey (Hg.): *Selbstorganisation in Wissenschaft und Technik, Wissenschaftsforschung Jahrbuch 2008*, Berlin: Gesellschaft für Wissenschaftsforschung, 107-119.

Merton, R. K. (1973), The Sociology of Science. Chicago: Chicago University Press.

Stirling, A. (2007). A general framework for analysing diversity in science, technology and society. *Journal of The Royal Society Interface* 4(15), 707–719.

White, H. C. (2008), Identity & Control. How Social Formations Emerge. Cambridge (MA): Harvard University Press.

White, H.C., Boorman, S. & Breiger, R. (1976), Social structure from multiple networks. I. Blockmodels of roles and positions. *American Journal of Sociology* 81, 730-779.

FULLERENE AND COLD FUSION: BIBLIOMETRIC DISCRIMINATION BETWEEN NORMAL AND PATHOLOGICAL SCIENCE

Marcus John[1] and Frank Fritsche[2]

[1] *Marcus.John@int.fraunhofer.de*
Fraunhofer Institute for Technological Trend Analysis INT, Appelsgarten 2, 53879 Euskirchen (Germany)

[2] *Frank.Fritsche@int.fraunhofer.de*
Fraunhofer Institute for Technological Trend Analysis INT, Appelsgarten 2, 53879 Euskirchen (Germany)

Introduction

Research activities on fullerenes were initiated by the discovery of this carbon allotrope by Kroto et al. (1985), which finally lead to a whole new research area. On the contrary, the fusion of hydrogen atoms at room temperature allegedly observed by Fleischmann & Pons (1989), proved to be false within a few years. Today the research on this phenomena, sometimes referred to as cold fusion, is generally considered as a prototypical example of pathological science.

This contribution presents a comprehensive comparison of these two scientific themes from a bibliometric point of view. We aim to look for patterns, which might allow to distinguish between normal and pathological science by means of bibliometric observables.

Method

To this end, we establish a bibliometric workflow, which consists of three different phases. The first phase comprises the formulation of a suitable search query, which aims to delineate the scientific theme of interest as accurately as possible. This step is of crucial importance for the workflow, since we want to include only those publications into the analyses, which are relevant for the specific scientific field.

Subsequently, the bibliometric data obtained with the ascertained search query are processed. The aim of this phase is to structure the data in a way, that they can be handled by an appropriate computer program and further information can be extracted, e.g. about the publishing countries. Furthermore, we try to correct potential errors (e.g. possible spelling errors within the address field).

These processed data are further analysed by calculating and visualizing a number of bibliometric observables and their time-dependent behaviour.

Considered bibliometric observables

For each of the two themes a number of bibliometric quantities are considered in order to elucidate the differences between them. In doing so we try to distinguish between normal and pathological science. Among others we analyse the following quantities:

1. The publication dynamics, viz. the number of papers published per year. In order to compare these data with the general trend of growing publication activities, we normalized them to a suitable reference year (see Figure 1 and 2).
2. The document types of the analysed publications.
3. The temporal evolution of the giant component's size of the co-author network (Bettencourt, Kaiser, & Kaur, 2009).

Each of these observables can be related to certain aspects of the process of scientific communication. While the publication dynamics obviously relates to the general interest in a scientific theme, the document types elucidate the scientist's preferred communication channel. Finally, the giant component reflects the process of formation of a scientific community.

Results

The results for the publication dynamics presented in Figure 1 (for the research on fullerenes) and Figure 2 (for the research on cold fusion) demonstrate, that these analyses allow a clear distinction between both themes. While the research on fullerenes follows after a short initial phase the general trend of growing publication activities, the research on cold fusion has almost disappeared from the scientific scene.

Intriguingly the analyses of the document types reveal considerable differences between the two themes. The researchers on fullerenes follow an expected pattern by mainly using articles and later on proceedings for their publication activities. On the contrary, the communication on cold fusion and the experiment of Fleischmann & Pons (1989) relies to a considerable extent on the publication of notes. This aspect might be interpreted as an indication of urgency or tentativeness in the presentation of the scientific results.

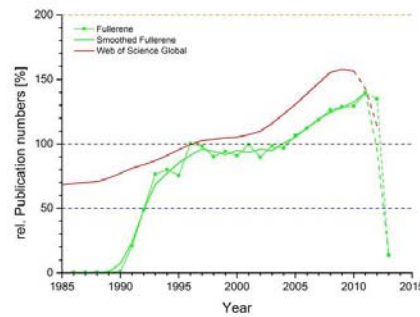


Figure 1. Relative publication numbers for the theme fullerene. The publication numbers for the research on fullerene and those for all publications within the data base are normalized to the year 1996, the year the Nobel Prize for chemistry was awarded to Kroto and colleagues.

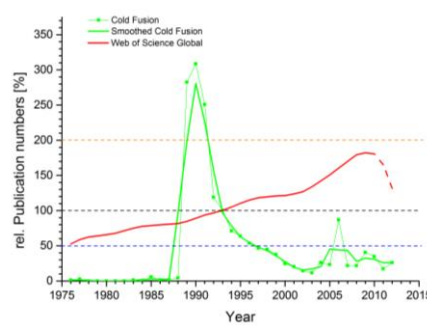


Figure 2. Relative publication numbers for the theme cold fusion. In this case the publication numbers have been normalized to the year 1993.

Finally, the results for the temporal evolution of the giant component of the co-author network reveal the most prominent difference between the two scientific themes. In case of the research on fullerenes a rather dense co-author network has emerged over time, where almost all authors are at least connected

indirectly. Thus, one clearly observes the formation of a scientific community. Contrarily, in case on the research on cold fusion, no such community has emerged and the co-author network remains rather disconnected.

Conclusion & Outlook

The aim of this contribution is a twofold. First of all we try to elucidate the differences between normal and (potentially) pathological science by bibliometric means. To this end, we combine the discussion of bibliometric observables with some small remarks on the scientific content of the analysed publications. It will be demonstrated, that a clear distinction between normal and pathological science seems to be possible from a bibliometric point of view. Furthermore, it will be discussed, whether or not it is possible to establish a kind of classification scheme for emerging topics, which might even serve as a kind of early-warning system. Of course such a scheme would have to be based on a larger number of considered themes.

References

Bettencourt, L. M. A., Kaiser, D. I., & Kaur, J. (2009). Scientific discovery and topological transitions in collaboration networks. *Journal of Informetrics, 3*(3), 210–221, from doi:10.1016/j.joi.2009.03.001.

Fleischmann, M., & Pons, S. (1989). Electrochemically induced nuclear fusion of deuterium. *Journal of Electroanalytical Chemistry and Interfacial Electrochemistry, 261*(2, Part 1), 301–308.

Kroto, H. W., Heath, J. R., O'Brien, S. C., Curl, R. F., & Smalley, R. E. (1985). C60: Buckminsterfullerene. *Nature, 318*, 162–163, from doi:10.1038/318162a0.

GEOGRAPHICAL ORIENTATION AND IMPACT OF FINLAND'S INTERNATIONAL CO-PUBLICATIONS

Reetta Muhonen[1], Hanna-Mari Puuska[2] and Yrjö Leino[3]

[1]*reetta.muhonen@uta.fi*
University of Tampere, Research Centre for Knowledge, Science, Technology and Innovation Studies, FI-33014 University of Tampere, (Finland)

[2]*hanna-mari.puuska@uta.fi*
University of Tampere, Research Centre for Knowledge, Science, Technology and Innovation Studies, FI-33014 University of Tampere, (Finland)
CSC – IT Center for Science P.O. Box 405, FI-02101 Espoo (Finland)

[3] *yrjo.leino@csc.fi*
CSC – IT Center for Science P.O. Box 405, FI-02101 Espoo (Finland)

Introduction

Several studies indicate that international co-publications have increased steadily in almost all countries. International co-publications are also cited more often than other publications. (e. g. Narin et al. 1991; Katz & Hicks 1997; Glänzel & Schubert 2001.) Similar results have been obtained in the Finnish context (NordForsk 2010, Muhonen et. al. 2012). This study extends the understanding of Finland's international research collaboration by scrutinizing its geographical orientation. We explore the trends and citation impact of Finland's co-publishing with various country groups by addressing the following questions:

1. How has international co-publishing developed in Finland in 1990–2009 with different country groups: Nordic countries, EU15+ countries[181], Baltic countries, Russia, other European countries[182], Africa, Asia, North America, Central and South America, and Australia and Oceania?

2. Have Finland's international co-publications with different country groups received more citations than Finnish publications on average?

Data and methods

The results presented in this study are based on the Thomson Reuters Web of

[181] EU15 states excluding the Nordic countries + Switzerland, that is: Austria, Belgium, France, Germany, Greece, Ireland, Italy, Luxembourg, Portugal, the Netherlands, Spain, Switzerland, and the United Kingdom.

[182] Albania, Andorra, Belarus, Bosnia-Herzegovina, Bulgaria, Czech Republic, Croatia, Cyprus, Hungary, Kosovo , Liechtenstein, Macedonia, Malta, Monaco, Moldova, Montenegro, Poland, Romania, San Marino, Serbia, Slovakia, Slovenia, Ukraine, the Vatican, and the former Yugoslavia.

Science database (WoS). Whole counting of publications is applied: if a publication includes authors from several country groups it is counted into each group. When calculating the citation scores, the publications are, however, fractionalized between countries.

Citation counts of Finnish publications with various country groups are compared to the world average by using the field normalized citation score. The number of citations cumulated by each publication is normalised to the average citation counts of all WoS publications in the relevant subject field which were published in that year and which represent the same WoS article type (Article, Review or Letter). Citation scores are calculated for publications in the latest studied period 2006–2008 (the year 2009 was excluded since its publications have not had time to cumulate an adequate amount of citations). We apply open citation window and exclude self-citations.

Table 1. Shares of country groups in Finland's international co-publications (WoS, 1990-2009).

	1990-1993	1994-1997	1998-2001	2002-2005	2006-2009
EU15+	38.0%	39.8%	46.7%	50.1%	53.8%
North America	35.4%	36.5%	32.8%	31.3%	30.3%
Nordic countries	23.6%	25.2%	25.5%	25.0%	26.1%
Other European	10.9%	10.8%	10.8%	9.9%	12.3%
Asia	7.8%	9.7%	10.1%	11.2%	14.3%
Russia	9.3%	9.5%	9.1%	7.8%	8.2%
Australia, Oceania	2.1%	2.5%	3.3%	3.5%	4.9%
Baltic countries	0.7%	3.2%	3.2%	3.3%	3.6%
South & Central America	2.2%	3.0%	2.3%	2.5%	2.6%
Africa	1.6%	1.3%	1.2%	1.3%	1.6%

Note. The sum of percentages of all groups exceeds 100 since a single publication can belong to several groups.

Results

The share of international co-publications in all Finland's WoS publications increased steadily from 25 percent 1990–1993 to almost one half (49%) in 2006–2009.

The co-authors of Finland's international co-publications most typically came from the EU15+ countries (Table 1). North America was the second and the Nordic countries the third most frequently co-publishing country group with Finland. No change took place in the order of the three most common country groups involved in co-publications, although their shares in Finland's international co-authorship have undergone different development patterns during 1990–2009.

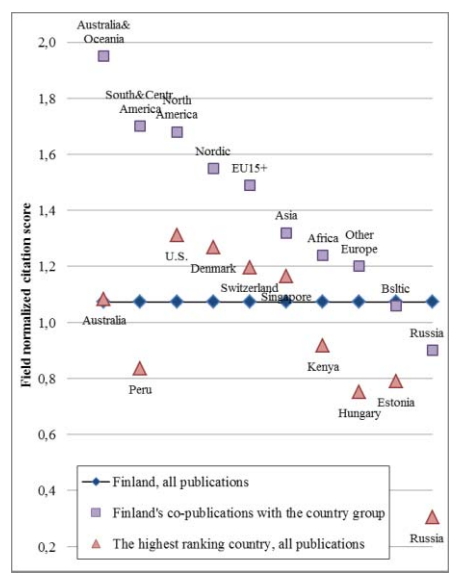


Figure 1. Field normalized citation score of 1) Finland's all publications 2) Finland's co-publications with the country groups, 3) all publications of the highest ranking country (i. e. the country with the highest citations score in the particular country group with minimum of 500 publications) (WoS, 2006-2008).

Figure 1 shows that the order of Finland's most impactful co-publishing country groups is well in line with the citation impact of the highest ranking countries in these groups. However, the highest citation score were displayed by Finland's co-publications with Australia & Oceania and South & Central America. They received 95 and 70 percent more citations than the world's publications on average. Exceptionally, the highest ranking countries in these country groups, Australia and especially Peru have much lower citation scores than corresponding countries in the other country groups.

Compared to all publications produced by Finnish scholars, the co-publications with all country groups are on average more frequently cited with one exception: The citation count of Finland's co-publications with Russia are lower than that of Finnish publications on average. Compared to Russia's own publications, however, these co-publications have cumulated considerably higher numbers of citations.

Discussion

The results of this study indicate that Finnish research has become more internationalized and the international collaboration benefits the impact of publications. Furthermore, collaboration with Finland has also been lucrative for other countries. The position of the United States as a hub of science is reflected as a high average citation score of North America's co-publications with Finland, but on the other way around, cooperation with Finland is also beneficial for the North America.

To be able to explain the high citation rates of the co-publications with Australia & Oceania and South & Central America a further analysis would be needed on the research fields in which Finland collaborates with different country groups. The further studies should also consider the possible variation in citation impact within the country groups.

References

Glänzel, W, Schubert, A. & Czerwon, H.-J. (1999). A bibliometric analysis of international scientific cooperation of the European Union (1985-1995). Scientometrics 45 (2), 185–202.

Katz, J. Sylvan & Hicks, Diana (1997). How much is a collaboration worth? A Calibrated bibliometric model. Scientometrics, 40 (3), 541–554.

Muhonen, R., Puuska, H.-M. & Leino, Y. (2012). International co-publishing in Finland. Helsinki: Reports of the Ministry of Education and Culture, Finland 2012:19.

Narin, F., Stevens, K., & Whitlow, E. S. (1991). Scientific cooperation in Europe and the citation of multinationally authored papers. Scientometrics, 21 (3), 313–323.

NordForsk (2010). Bibliometric research performance indicators for the Nordic countries. A Publication from the NORIA-net "The Use of bibliometrics in research policy and evaluation activities".

GLOBAL RESEARCH STATUS IN LEADING NUCLEAR SCIENCE AND TECHNOLOGY JOURNALS DURING 2001–2010: A BIBLIOMETRIC ANALYSIS BASED ON ISI WEB OF SCIENCE

Amir Hossein Mardani[1] and Fereshteh Didegah[2] and Shahram Abdiazar[3]

[1] *Mardani3@gmail.com*
Tehran University of Medical Sciences, School of Medicine, Ghods Str, Keshavarz Blv ,Tehran (Iran)

[2] *fdidgah@gmail.com*
Statistical Cybermetrics Research Group, School of Technology & Research Institute for Information and Langauge Processing, University of Wolverhampton (UK)

3 Sh.abdiazar@gmail.com
University of Maraghe, Madar Sq, Amirkabir Blv, Maraghe (Iran)

Introduction

The present study will provide an assessment of global orientations toward various disciplines in terms of Nuclear Science and Technology (NST) research. This study will try not to content itself only with a preliminary bibliometric performance assessment of individuals, institutions, and countries and will attempt to explain the developments in this field beyond the observed patterns. Therefore, we tried to use advanced bibliometric measures such as filed citation scores and analytical methods such as network analysis. At the same time, we measured a considerable amount of publications associated with NST and offered substantive arguments for observed bibliometric patterns so as to provide potential directions for future research.

Material and methods

The records used in this study are based on bibliographic data retrieved from the WOS database. First, 35 journals were identified which were listed under the subject category of NST in the database of JCR in 2011. Next, we retrieved all the publications of this journal which were indexed in the WOS during 2001 to 2012. All types of collaborations were categorized based on the affiliation addresses of the authors: "single country publication", "internationally collaborative publication", "single institute publication" , and "inter-institutionally collaborative publication". Furthermore, the NetDraw software was used to illustrate the international collaboration network (Borgatti, 2006). To determine performance of papers across fields, in terms of assessing impact of the papers across fields, we calculated the field normalized measured impact ratios (CPP/FCSm) of NST publications for each field.

Results

In general, 85198 records were retrieved from WOS during 2001 to 2010. Although the annual number of records

during the period has rather increased, as can be seen in figure 1, this increase has not been very conspicuous and the research output does not reveal a significant growth in NST scientific productions. It even has experienced a downfall in some years such as 2008 and 2010. The exponential regression test supported that these publications have a 1.3% growth rate. The results were accurate and reliable at a confidence level of 95 percent (significance= 0.001).

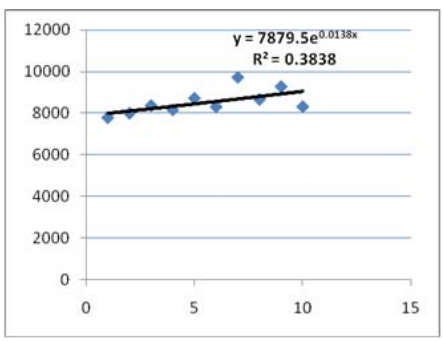


Fig 1. The growth rate of publications

The NST publications fall into 18 ISI recognized subject categories. Figure 2 provides a spectral analysis of the research output of NST across those fields. Results revealed that of the 18 fields the largest impact (CPP/FCSm = 4.2) was for Nuclear Physics, but still substantial impact level (CPP/FCSm above 1.5) was also observed and included the fields of Nuclear Science & Technology, Spectroscopy, Analytical Chemistry, multidisciplinary sciences, Biology, and Geosciences. The field normalized citation scores indicated that publications in these fields were highly influential and visible, far exceeding the number of citations expected for publications of the same time period (Rosas et al. 2012; Liu et al. 2012; Van Raan, 2008).

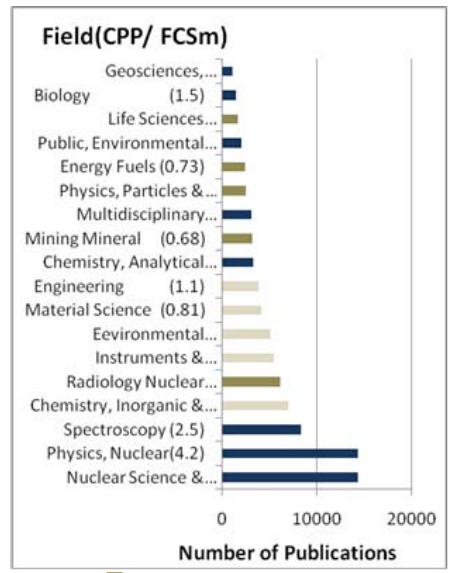


Fig 2. Research output profile of NST across the top 18 fields, 2001–2010

Using the NetDraw software, we were able to illustrate the international collaboration network which consists of 25 countries (Figure 3). As can be observed in Figure 3, the USA is at the center of the NST research cluster which indicates that the USA plays an important role in the network. The USA is in the central part of the international collaboration network; hence, it is the main partner for many research prolific countries such as Germany, Japan, and Italy. However, the network suggests the strongest relationship between the USA and Germany (with 1174 counts of co-authored records).

We recorded the publication indicators for twenty five countries which were prolific in NST scientific productions. The data indicate geographical inequalities in the publications. From the 82738 remaining records, 61753 one of them (74.6%) were published independently and without any collaboration from another country.

20985 records (25.4%) were published with international collaboration of more than one country.

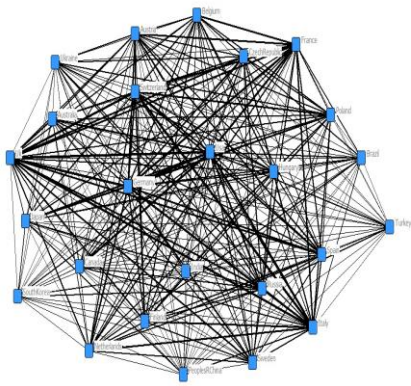


Fig 3. International collaboration network of 25 most central countries in Nuclear Science & Technology research

The USA comes first in the rankings of countries that are prolific in NST research. Accounting for one fifth of scientific productions, the USA also ranks first in internationally collaborative publications (20% in all) and as a single country (22% in all). After that, Japan has published the highest number of records (11471). Germany (10119), France (6793), Italy (6561), and Russia (5796) come right afterward. Even though internationally collaborative publications comprise 23% of all the publications in the USA, they are exceeded by in some countries. Although ranking sixth in the number of publications, Russia contributes to 41% of internationally collaborative publications. Therefore, based on the collaboration pattern, it can be implied that some countries like Russia tend to collaborate with researchers from other countries. It is while most countries like Ukraine are more limited with regard to international collaborations than domestic ones.

References

Borgatti, Steve. (2006). Netdraw: Network Visualization. Analytic Technologies, Inc.

Liu, X., Zhan, F., Hong; S., Niu; B. &Liu, Y. (2012). A bibliometric study of earthquake research: 1900–2010. Scientometrics. DOI 10.1007/s11192-011-0599-z

Rosas ,S., Kagan .J., Schouten, J., Slack, P.& Trochim, W. (2011) Evaluating Research and Impact: A Bibliometric Analysis of Research by the NIH/NIAID HIV/AIDS Clinical Trials Networks. PLoS ONE, 6(3).

Van Raan , A. (2008). R & D evaluation at the beginning of a new century. Research Evaluation, 8, 81–6.

GROUPS OF HIGHLY CITED PUBLICATIONS: STABILITY IN CONTENT WITH CITATION WINDOW LENGTH

Nadine Rons[1]

[1] *Nadine.Rons@vub.ac.be*
Vrije Universiteit Brussel, Research Coordination Unit and Centre for R&D Monitoring (ECOOM), Pleinlaan 2, B-1050 Brussels (Belgium)

Introduction

The growing focus in research policy worldwide on top scientists makes it increasingly important to define adequate supporting measures to help identify excellent scientists. Highly cited publications have since long been associated to research excellence. At the same time, the analysis of the high-end of citation distributions still is a challenging topic in evaluative bibliometrics. Evaluations typically require indicators that generate sufficiently stable results when applied to recent publication records of limited size. Highly cited publications have been identified using two techniques in particular: pre-set percentiles, and the parameter free Characteristic Scores and Scales (CSS) (Glänzel & Schubert, 1988). The stability required in assessments of relatively small publication records, concerns size as well as content of groups of highly cited publications. Influencing factors include domain delineation and citation window length. Stability in size is evident for the pre-set percentiles, and has been demonstrated for the CSS-methodology beyond an initial citation period of about three years (Glänzel, 2007). Stability in content is less straightforward, considering for instance that more highly cited publications can have a later citation peak, as observed by Abt (1981) for astronomical papers. This paper investigates the stability in content of groups of highly cited publications, i.e. the extent to which individual publications enter and leave the group as the citation window is enlarged.

Data and methodology

Database, document type and time frame

The bibliometric data were obtained from the online Web of Science. This study focuses on articles as primary vehicles for new research results (reviews, less frequent and typically more highly cited, are in itself related to high esteem). Results were calculated for articles published in 2004. Citations were collected up to 7 years after publication, until 2011.

Aggregation level

The group of top-cited publications is highly dependent on the level of aggregation at which these are determined (Zitt, Ramanana-Rahary & Bassecoulard, 2005). In this paper, the reference sets within which highly cited publications are identified are the partition cells formed by the structure of overlapping subject categories (Rons, 2012). These partition cells are an aggregation level of intermediate size, situated between journals and entire subject categories, proposed particularly

for usage at the level of individual scientists.

Domains

Data were collected from a domain with fast citation characteristics (a sub-domain of physics), and from a domain with slow citation characteristics (mathematics). In both domains, three adjacent partition cells are observed, containing all journals assigned to a particular combination of subject categories:

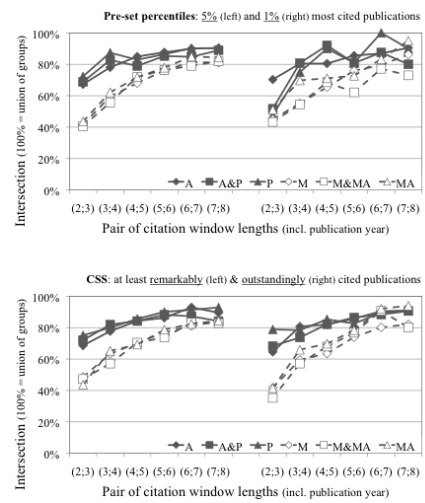
- 'Astronomy Astrophysics' only (A), 'Physics Particles Fields' only (P), and both 'Astronomy Astrophysics' and 'Physics Particles Fields' (A&P), with 8047, 1977 and 2440 articles respectively.
- 'Mathematics' only (M), 'Mathematics Applied' only (MA), and both 'Mathematics' and 'Mathematics Applied' (M&MA), with 10022, 3938 and 3286 articles respectively.

Groups of highly cited publications

Groups of highly cited publications were identified using both the technique of pre-set percentiles (5% and 1%), and the CSS-methodology (at least 'remarkably' and 'outstandingly' cited publications, stabilizing in size towards 8 to 11%, and 2 to 3% respectively in the largest citation windows).

Results and discussion

The citation distributions vary with domain and highly cited level. Peak citation years tend to fall later for more highly cited publications. Differences depend on the domain, with in the physics cells 0 to 3 years between the top-1% or outstandingly cited and the less cited publications, and in the mathematics cells 1 tot 2 years between the top-5% or at least remarkably cited and the less cited publications. In particular for publications below the top-1% and outstandingly cited, peak citation years in the slow mathematics domain fall later than in the fast physics domain (3 to 6, and 1 to 2 years after the publication year respectively). Figure 1 shows how groups of highly cited publications converge in content with increasing citation window length, in a similar way for both methodologies used. A yearly fluctuation in content remains between the larger consecutive citation windows.



Data sourced from Thomson Reuters Web of Knowledge (formerly referred to as ISI Web of Science). Web of Science (WoS) accessed online 10-11.01.2013 and 14.02.2013.

Figure 1. Intersection of groups of highly cited publications for consecutive citation window lengths.

Convergence in content varies with domain in particular, and also with highly cited level. From a citation window of 3 years, groups of highly cited publications identified in consecutive citation windows have a majority of publications in common. This finding is in accordance with citation distribution models reflecting the conjecture that for most papers the initial pulse of citations determines its future citation history (Price, 1976). To have at least 80% of highly cited

publications in common with the subsequent citation window requires 3 to 4 years for the physics cells, and 5 or more years for the mathematics cells. Changes in status for 1 out of 5 highly cited publications over the years might however still significantly influence a comparison of publication records of individual scientists. In practice citations are often collected in about the same period as the observed publications, for instance extended with one year. The results shown in Figure 1 indicate that in particular the potential error induced by a citation window as short as two years would need to be considered in the design of an approach for the identification of highly cited publications.

Conclusions

The results of the present study show how, in stable size groups of highly cited publications, the content of individual publications converges with an extending citation window. To the author's knowledge, no similar investigations of the extent to which individual publications enter and leave the group of highly cited publications as the citation window is enlarged, figure in scientific literature. The process evolves towards a limited remaining yearly fluctuation and depends on domain and highly cited level. Additional factors outside the scope of this paper may also exert an influence, such as the structure within which the highly cited publications are identified, including subject category as well as publication based structures. Stability in content of groups of highly cited publications is relevant in particular in a context of publication records of limited size, such as those of individual scientists. The studied domains with different bibliometric characteristics indicate that a same stability in content can require citation windows of different lengths depending on the domain. Whether a sufficient stability in content can be attained for successful general and comparative applications at the micro-level, with citation windows of acceptable size in an evaluative context, requires further investigation.

Acknowledgments

This paper is related to research on individual excellence carried out at the Flemish Centre for R&D Monitoring (ECOOM).

References

Abt, H.A. (1981). Long-term citation histories of astronomical papers. *Publications of the Astronomical Society of the Pacific*, 93(552), 207-210.

Glänzel, W. (2007). Characteristic scores and scales: A bibliometric analysis of subject characteristics based on long-term citation observation. *Journal of Informetrics*, 1(1), 92-102.

Glänzel, W. & Schubert, A. (1988). Characteristic scores and scales in assessing citation impact. *Journal of Information Science*, 14(2), 123-127.

Price, D. de S. (1976). A general theory of bibliometric and other cumulative advantage processes. *Journal of the American Society for Information Science*, 27(5-6), 292-306.

Rons, N. (2012). Partition-based field normalization: An approach to highly specialized publication records. *Journal of Informetrics,* 6(1), 1-10.

Zitt, M., Ramanana-Rahary, S. & Bassecoulard, E. (2005). Relativity of citation performance and excellence measures: from cross-field to cross-scale effects of field-normalization. *Scientometrics*, 63(2), 373-401.

HEAPS' LAW: A DYNAMIC PERSPECTIVE FROM SIMON'S MODEL

YiFei Zhang[1], Shi Shan[2], JunPing Qiu[3] and ChunNing Yan[1]

[1]*zhangyifei@shu.edu.cn*
Shanghai University, School of Management, 99 Shangda Road, 200444 Shanghai (China)

[2]*shanshihill@126.com*
Shanghai University, Center for Information Studies, 99 Shangda Road, 200444 Shanghai (China)

[3]*jpqiu@whu.edu.cn*
Wuhan University, Research Center for Chinia Science Evaluation, Luojia Hill, 430072 Wuhan (China)

Introduction

This papers focuses on the growth of vocabulary in actual texts. Here we examine how the macroscopic text properties, like word occurrence and text densification change as the text grows. This work had influence on thinking about fundamental structure properties of texts varying with their length. For example, to date, it was commonly believed that the average word occurrence of texts remains constant as the number of texts increases. In fact texts densify with the number of texts as the total number of words in m texts $K(m)$ increasing as $K(m) \propto N(m)^{1+\varepsilon}\,(\varepsilon > 0)$ with the total number of different words in m texts $N(m)$, where $\varepsilon > 0$. We use $e(m)$ to denote the average word occurrence in m texts, then $e(m) = \frac{K(m)}{N(m)} \propto N(m)^{\varepsilon}$.

Obviously, the average word occurrence grows with the number of samples (i.e., texts). Two models describing the growth of vocabulary are worth mentioning. The first one, a stochastic process put forward by Simon (Simon, 1955), simulates the dynamics of text generation as multiplicative process that leads to power law distributions of word frequencies. The second model is due to Heaps (Heaps, 1978), who developed a power-law relation between the text length k and the number $n(k)$ of different words in a text, i.e., $n(k) = ck^{1-\theta}\,(c > 0,\ \theta \in (0,1))$, which is usually referred to as Heaps' law.

Simon's model intend to capture the essential features of real text generation by specifying how words are added to a text, as follows. The probability that the $(k+1)$st word added to a text will be a word that has already occurred i times $(i \geq 1)$ is proportional to $if(i,k)$, where $f(i,k)$ is the number of distinct words (types) that have occurred exactly i times each in the first k words of text. The probability that the $(k+1)$st word added to a text will be a word that has not occurred before is $\rho \in (0,1)$. Simon developed the following model.

$$f(1,k+1) - f(1,k) = \rho - \frac{1-\rho}{k} f(1,k)$$

and

$$f(i,k+1)-f(i,k)=\frac{1-\rho}{k}((i-1)f(i-1,k)-if(i,k))$$
$$(i=2,3,\cdots,k+1).$$

In the Simon model, the rate at which new words appear is a constant, such that the vocabulary size $N(k)$ for a text containing k words, on the average, is $N(k)=\rho k$. Obviously, $\Delta N(k)=N(k+1)-N(k)=\rho$, which is the rate at which new words enter. Heaps' law shows that the growth of vocabulary in actual texts is typically sublinear, i.e., $N(k)=ck^{1-\theta}$ with $c>0$ and $\theta\in(0,1)$, and

$$\Delta N(k)=N(k+1)-N(k)=(1-\theta)ck^{-\theta}(1+o(1)).$$

The latter means that the rate at which new words appear is not a constant but a gradually decreasing function of k.

An Extension of the Simon Model

We will consider an extension of the Simon model.

Let k be the text length. $f(i,k)$ is the expected number of distinct words that occur i times each in a text with length k. $N(k)=\sum_{i=i}^{k}f(i,k)$. $\rho(k)$ is the rate at which new words enter. $M(k)=\sum_{i=1}^{k}\rho(i)$. $N_+=\{1,2,\cdots\}$. A positive function $\varphi(k)$ on N_+ varies gradually, if and only if $\frac{\varphi(k+1)}{\varphi(k)}=1+o(\frac{1}{k})$. $p(i)=\lim_{k\to\infty}f(i,k)/\sum_{j=1}^{k}f(j,k)$ is called the steady-state distribution, if the limit exists. $\Gamma(\cdot)$ is the usual gamma function. We use "~" to mean that the ratio of the quantities on either side of the symbol converges to one.

We extent the Simon model in the following way.

$$f(1,k+1)-f(1,k)=\rho(k+1)-K(k)f(1,k),$$
$$f(i,k+1)-f(i,k)=K(k)((i-1)f(i-1,k)-if(i,k))$$
$$,(i=2,3,\cdots,k+1),$$

where $1-\rho(k+1)=\sum_{i=1}^{k}K(k)if(i,k)$.

Next, we assume that

$$\sum_{i=1}^{k}if(i,k)\sim k.$$

From this assumption, it follows that

$$f(1,k+1)-f(1,k)=\rho(k+1)-\frac{1-\rho(k+1)}{k(1+o(1))}f(1,k) \quad (2.1)$$

$$f(i,k+1)-f(i,k)=\frac{1-\rho(k+1)}{k(1+o(1))}((i-1)f(i-1,k)-if(i,k))$$
$$(i=2,3,\cdots,k+1), \quad (2.2)$$

We obtain the following results.

Theorem 1 Suppose that (2.1) and (2.2) hold and $\rho(k+1)\to\rho\in(0,1)$ (as $k\to\infty$), then

(1) $\lim_{k\to\infty}\frac{f(i,k)}{\sum_{j=1}^{k}f(j,k)}=\frac{\alpha\Gamma(\alpha+1)\Gamma(k)}{\Gamma(k+\alpha+1)}$,

where $\alpha=\frac{1}{1-\rho}>1$, and

(2) $N(k)\sim\rho k$.

Theorem 2 Suppose that (2.1) and (2.2) hold, $\rho(k)\to 0$ (as $k\to\infty$),

$$\lim_{k\to\infty}\frac{k\rho(k+1)}{M(k)}=\beta\in(0,1)$$

and

$$\lim_{k\to\infty}\frac{f(i,k)}{M(k)}=q(i)\,(i\in N_+),$$

then

(1) $\lim_{k\to\infty}\frac{f(i,k)}{\sum_{j=1}^{k}f(j,k)}=\frac{\beta\Gamma(\beta+1)\Gamma(k)}{\Gamma(k+\beta+1)}=q(i)$

with $\beta\in(0,1)$, and

(2) $N(k)\sim\varphi(k)k^{\beta}$ $(\beta\in(0,1))$, where $\varphi(k)$ varies gradually.

Remark Theorem 2 shows that (under some conditions) when the rate at which new words enter gradually approaches zero a general form of Heaps' law will emerge.

References

Egghe, L. (2007). Untangling Herdan's law and Heaps' law: mathematical and informetric arguments. *Journal of the American Society for Information Science and Technology*, 58, 702-709.

Heaps, H. S. (1978). Information Retrieval, Computational and Theoretical Aspects. New York: Academic Press.

Lansey, J. C. & Bukiet, B. (2009). Internet search result probabilities: Heaps' law and word associativity. *Journal of Quantitative Linguistics*, 16, 40-66.

Leijenhorst, D. & Weide, T. (2005). A formal derivation of Heaps' law. *Information Science*, 170, 263-272.

Lu, L., Zhang, Z.-K. & Zhou, T. (2010). Zipf's Law Leads to Heaps' Law: Analyzing Their Relation in Finite-Size Systems. *PLoS ONE*, 5, e14139.

Shi, D. H. (2011). *Theory of Network Degree Distributions* (in Chinese). Beijing: Higher Education Press.

Simon, H. (1955). On a class of skew distribution functions. *Biometrika*, 42, 425-440.

Zanette, D. & Montemurro M. (2005). Dynamics of text generation with realistic Zipf's distribution. *Journal of Quantitative Linguistics*, 12, 29-40.

Zhang, Z.-K. et al. (2008). Empirical analysis on a keyword-based semantic system. *European Physical Journal B*, 66, 557-561.

Zipf, G. K. (1936). The Psycho-Biology of Language. An Introduction to Dynamic Philology. London: Routledge.

HOW EFFECTIVE IS THE KNOWLEDGE TRANSFER OF A PUBLIC RESEARCH ORGANIZATION (PRO)? FIRST EMPIRICAL EVIDENCE FROM THE SPANISH NATIONAL RESEARCH COUNCIL

Elba Mauleón[1], Gian Carlo Cainarca[2] and Cinzia Daraio[3]

[1]*elba114@hotmail.com*
Postdoctoral Fellowship supported by the Spanish Ministry of Education through *FECYT*

[2] *cainarca@dist.unige.it*
DIST, University of Genoa (Italy)

[3]*daraio@dis.uniroma1.it*
Department of Computer, Control and Management Engineering Antonio Ruberti
University of Rome La Sapienza, Rome (Italy)

Introduction

Scientific research is fundamental for countries to progress, but increasingly requires greater economic investment (Van Raan, 1996). Thus, some issues of concern for governments are raising funds through international calls, the proper distribution of funds, and the establishment of procedures for assessing the activity of groups and research institutes, to ensure proper use of limited resources (Bornmann, Mutz and Daniel; 2008). In this sense, to achieve the efficiency in the use of these resources is a goal of science policy, and the improvement of the instruments that measure the efficiency in organizations is a staple.

Objectives

In this paper we analyse the efficiency of knowledge production of a PRO and we investigate also on the much less explored issue of the "effectiveness" of the knowledge production, i.e. the knowledge transfer activity and its impact at territorial level. At this purpose, the Spanish National Research Council (Consejo Superior de Investigaciones Cientificas - CSIC) offers a unique opportunity to analyse a complex state-owned, multisectoral, multidisciplinary, public research body, articulated into different research areas spread at territorial level all over Spain.
The process of knowledge transfer may be articulated in:

a) *technology transfer in the narrow sense*: based on codified knowledge and measured by quantitative indicators (such as patents and so on) that impacts on firms and the innovation system of the economy; and
b) *technology transfer in the broad sense*: based on non codified knowledge more difficult to measure, that affects the general public and the society.

So, we provide a comparative analysis of CSIC research institutes to investigate

on their efficiency in the production of knowledge, and to analyse the effectiveness of their production i.e. how the knowledge produced is transferred to the society.

Methodology

As a result of the transformation of the Spanish National Research Council (CSIC) into a state agency, strategic plans (Plan de Actuación, 2006-2009) have been implemented to prepare the institution's activities and establish a set of indicators to monitor its activity. The implementation of this action plan has generated a wealth of qualitative and quantitative information and data has been provided by the centres and institutes on how they operate. We used the following data, which is available for every centre and institute:

a) Human resources: the number of researchers.
b) Budget
c) Funding obtained through contracts and agreements signed with R & D companies, with the public sector (ministries or agencies, regional governments,) and non-profit institutions.
d) Funds allocated to research: funded research projects are considered by source (competitive national and regional calls and EU Framework Programme call).
e) Number of international scientific publications.
f) Impact factor of research journals.
g) Training activities: including the number of theses that have been presented, and the participation of researchers in master's and doctoral courses.

To consider the importance and influence of the environment (economic, social and political) the following variables relative o the region in which the institutes are located are included in the study: a) GDP per capita= Gross Domestic Product per capita (Euros); b) R&D Intensity= Research & Development expenditure as % gross domestic product; c) BERD= % of R&D expenditure financed by the business enterprise sector.; e) Patent/inhabitant: number of patents per million inhabitants (due to the small number of annual patents in some regions, a three-year average (2002-2004) is considered to reduce year-to-year variability).

We assess the efficiency of CSIC as a whole measuring the relative efficiency of each research institutes by using a Data Envelopment Analysis approach (DEA, Farrell, 1957, Charnes Cooper and Rhodes, 1978). After that we study the determinants of the efficiency of the CSIC institutes by using a two stage bootstrap based approach in nonparametric efficiency analysis (Simar and Wilson, 2007).

Table 1. Distribution of centres analyzed by area.

Research area	N. centres	% centres analyzed
Biology and Biomedicine	21	76.19
Natural Resources	19	73.68
Physical Sciences and Tech.	24	66.67
Material Science and Tech.	9	100.00
Chemical Sciences and Tech.	14	64.29
Food Science	6	50.00
Agriculture Science	12	83.33

Results

The CSIC is the main research institution in Spain. It is multidisciplinary and it is organized into eight scientific areas. The CSIC employs around 2,200 researchers (3% of human resources related to research in Spain) and contributes 20% of Spanish output to the international database Web of Science, just behind

Universities (63%) and the Health Sector (23%) (Gómez et al, 2010). The table 1 shows the scientific areas and brief summary about the number of centres/institutes analyzed in each research area.

In particular we will analyse partial models of knowledge production, including only scientific production (i.e. international and national production) and a comprehensive model including both technological (i.e. patents) and scientific activities.

Table 2 illustrates the specifications of the models that we empirically estimate in this first empirical effort.

Inputs specification	*Outputs specification*
MODEL 1-Scientific production	
- Total scientists	- N. of ISI papers
- Post-doctoral researchers	- N. of non ISI papers
- Funds from research projects	- Other publications
MODEL 2-Technological and Scientif Production	
- Total scientists	- Number of patents
- Post-doctoral researchers	- N. of ISI papers
- Funds from research projects	- N. of non ISI papers
	- Other publications
MODEL 3-Knowledge transfer	
- Gross Domestic Product	- N. of PhD students
- Agglomeration Index	- N. of post graduate courses
- Efficiency level of knowledge production	- N. of contracts with third parties

We will then model how knowledge transfer activities are carried out by CSIC institutes.

In modelling knowledge transfer we consider the efficiency of knowledge production as one input that contributes to the knowledge transfer activity. Of course this input has to be combined with the other inputs that affect the knowledge transfer, that we proxy with the Gross Domestic Product (GDP) and the Agglomeration Index. We use the GDP as an input to proxy the lively and richness of the regional context in terms of knowledge activities and knowledge transfer channels (both formal and informal ones); see at this purpose, Bishop, D'Este and Neely (2011). Our choice of the Agglomeration Index as an input is consistent with previous literature that found the existence of agglomeration economies - related to labor market pooling, input sharing and knowledge spillovers- that attenuate with distance (see e.g. Rosenthal and Strange, 2004).

Discussion

In this research we provide some first preliminary investigation on issues that are not very explored in the literature, related to the evaluation of the effectiveness of knowledge transfer that generates from the knowledge produced by the Spanish CSIC. From the first investigations we conducted the following policy implications seem to emerge:

1. The evaluation of Public Research Organizations should relate not only the formal scientific and technological activities carried out (which relays on publications and patents) but should also be based on teaching, contracts with third entities and in general other third mission activity indicators;
2. The localization of research institutes should be planned not only in agglomerated and concentrated big areas with other research centres but should take also into account the specificities of the territory and the match between research activities carried out and territorial devotion.

Further investigations will be directed to check if the preliminary results we found in this paper are consistent by applying the recent developed nonparametric conditional methodology (Daraio and Simar, 2007; Daraio, Simar

and Wilson, 2010; Badin, Daraio and Simar, 2012).

Acknowledgments

This study has been carried out by Elba Mauleón during her Postdoctoral Program supported by the Spanish Education Ministry and FECYT.

References

Daraio, C., Simar L. (2007), Advanced Robust and Nonparametric Methods in Efficiency Analysis. Methodology and Applications, Springer, New York.

Badin, L., Daraio, C. and L. Simar (2012), How to measure the impact of environmental factors in a nonparametric production model, *European Journal of Operational Research*.

Charnes, A., Cooper, W., and Rhodes, E. (1978), Measuring the efficiency of decision making units, European Journal of Operational Research, Volume 2, pp. 429-444.

Farrell, M.J. (1957), The measurement of productive efficiency, Journal of the Royal Statistical Society, Series A 1, 20, 253-281.

Gómez, I.; Bordons, M.; Morillo, F.; González-albo, B.; Candelario, A.; Herrero, A. (2010). La actividad científica del CSIC a través del Web of Science. Estudio bibliométrico del periodo 2004-2008. Madrid, IEDCYT, CCHS, CSIC, 2010. 1.039 pág. http://hdl.handle.net/10261/22941

Plan de Actuación, 2006-2009. (2006). Consejo Superior de Investigaciones Científicas.

Simar, L and P. W. Wilson (2007), Estimation and Inference in Two-Stage, Semi-Parametric Models of Production Processes, Journal of Econometrics, 136(1), 3164.

HOW MUCH MATHEMATICS IS IN THE *BIG TWO* AND WHERE IS IT LOCATED?

Gunar Maiwald[1]

[1]*maiwald@zib,de*
Konrad-Zuse-Zentrum für Informationstechnik Berlin (ZIB), Takustrasse 7,
14195 Berlin (Germany)

Introduction

Scopus and Web of Science (WoS) are the two major and relevant citation databases. Both are multidisciplinary and base on own journal-based classification systems. The variety of journals in both databases depends on different selection criteria. Therefore not all scientific journals of a specific discipline are indexed. In recent studies the coverage of different disciplines was investigated, e.g. Oncology (López-Illescas, De Moya-Anegón & Moed, 2008) and Library and Information Science (Abrizah, Zainab, Kiran & Raj, 2012). The present paper addresses some issues of Mathematics and its representation and classification in WoS and Scopus.

Mathematics is a broad subject which ranges from Pure Mathematics such as Algebra, Analysis or Topology to Applied Mathematics such as Computing, Numerical Analysis and Mathematical Physics. This spectrum is reflected by the Mathematical Classification System (MSC). The recent version from 2010 lists more than 5000 classes, hierarchically nested within 63 top-level classes. Compared to MSC, the categories of both databases are less specific. Scopus lists 15 explicit Mathematics classes. WoS does without, but other studies assume five categories (Bensman, Smolinsky & Pudovkin, 2010)

Mathematics comes with two major reviewing databases, Mathematical Reviews and Zentralblatt MATH (ZB). Both declare its deep coverage of Mathematics literature and each can be seen as comprehensive evidence of Mathematics literature. According to its policies not all subject-related content is indexed by Scopus and WoS.

This paper addresses the following questions: (a) To what extent are Mathematics related journals indexed in WoS and Scopus? (b) Which subject categories in Scopus and WoS are dominated by Mathematics journals? (c) How do WoS and Scopus subject categories relate to MSC?

Methods

In order to answer the present questions, the following methods refer to the order of questions above. (a) The analysis bases upon the journals indexed in ZB. This list is available with a web interface (FIZ Karlsruhe GmbH, 2012). Records without ISSN were eliminated, and those with identical ISSN were aggregated to a single record. The intersection between ZB, Scopus and WoS was determined by pairwise comparison of ISSN. (b) The result was breaken down to the subject categories of Scopus and WoS. Per each category the fraction of journals indexed in ZB was determined to all journals of this category. (c) In contrast to WoS and Scopus the MSC in ZB is assigned per

article. Articles have up to five classes assigned. To derive a relation between MSC and Scopus- and WoS-categories a pairwise matching of journal articles was done. Because literature indexed in Scopus is limited to recent articles, the considered period is 1996-2011. In further processing the 63 top-level classes was taken in focus. To determine the intersection efficiently, all datasets were partitioned into subsets of ISSN and year. Therefore the pairwise comparison of two datasets is limited to the pairwise comparison of corresponding subsets. Two articles in corresponding subsets are considered to be equal if normalized titles are equal or relevant metadata (volume, issue, first page, last page) match. Per each category in WoS and Scopus the fraction of each top-level MSC was calculated by pairwise article-matching.

Results

The results presented refer to the order of questions above. (a) At the time of inquiry were 3.841 journals indexed in ZB, 27.630 in Scopus and 18.247 in WoS. About 43% of journals indexed in ZB could be detected in WoS or Scopus. From those detected ones two-third are listed in both databases. Looking at the remaining journals, again

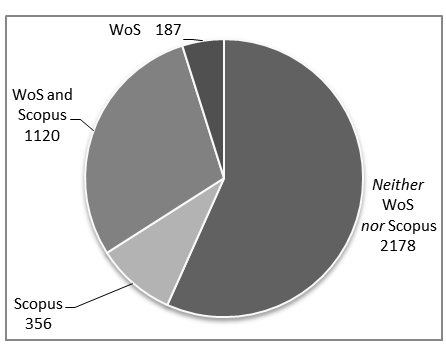


Figure 1. Journals from Zentralblatt MATH indexed in Scopus and Web of Science

two-third are indexed in Scopus and one-third in WoS (Figure 1).
(b) The journals indexed in ZB are assigned to nearly all categories in Scopus and WoS, but as expected, with a skewed distribution (Table 1). Fourteen subject categories of Scopus have a share of 60% and more of Mathematics journals, among them a category for publications in the field of Physics (*Statistical and Nonlinear Physics)*. At the same level there are six WoS-categories, two among them (*Logic, Statistics & Probability*) are not identified as Mathematics categories in other studies (Bensman et al., 2010).

Fraction of Mathematics journals	Number of categories in Scopus	Number of categories in WoS
>= 80%	7	5
>= 60%	14	6
>= 40%	21	16
>= 20%	30	31

Table 1. Number of Subject Categories in Scopus and Web of Science and their fraction of Mathematics journals

(c) As a basis for the analysis 1.1 Mio articles indexed in ZB and published between 1996 and 2011 were drawn from sample. About two-third of them could match with articles in Scopus or WoS, and more than 50% matched with records in both databases. The subject categories in Scopus and WoS show different behaviour to MSC-classes.
The big categories in Scopus (*Applied Mathematics, Mathematics (all)*) are spread over a big number of MSC-classes. Some of the narrower categories (*Statistics and Probability, Logic*) are dominated by only few MSC-classes (Figure 2).
Similar conduct can be seen in WoS: the big categories are covered almost evenly by a big number of MSC-classes. Some

of the narrower categories are located mainly in only few MSC-classes.

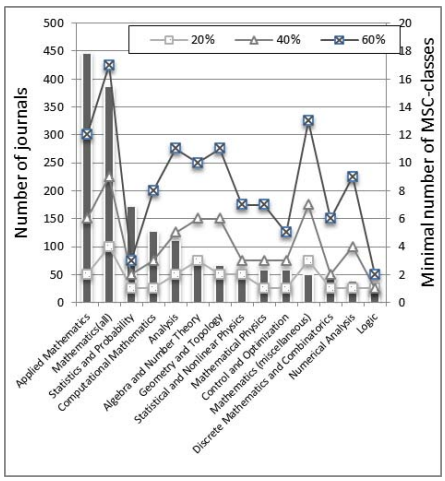


Figure 2. Minimal number of MSC-classes that cover 20%, 40% and 60% of a subject category – the 14 Scopus-Categories with a fraction of at least 60% of Mathematics journals

Discussion

Future research might investigate dynamic aspects between MSC and the subject categories in Scopus and WoS.

Acknowledgments

Certain data included herein is derived from the "Science Citation Index Expanded (SCIE) - Tagged Data" prepared by Thomson Reuters (Scientific) Inc. (TR®), Philadelphia, Pennsylvania, USA: © Copyright Thomson Reuters (Scientific) 2012, the Scopus® Custom Data datasets, Elsevier B.V., Amsterdam, The Netherlands and the "Zentralblatt MATH", © FIZ Karlsruhe GmbH, 2012. All rights reserved.

References

Abrizah, A., Zainab, A. N., Kiran, K., & Raj, R. G. (2012). LIS journals scientific impact and subject categorization: a comparison between Web of Science and Scopus. *Scientometrics*, *94*(2), 721–740. doi:10.1007/s11192-012-0813-7

Bensman, S. J., Smolinsky, L. J., & Pudovkin, A. I. (2010). Mean Citation Rate per Article in Mathematics Journals: Differences From the Scientific Model. *World*, *61*(7), 1440–1463. doi:10.1002/asi.21332

FIZ Karlsruhe GmbH. (2012). Zentralblatt MATH - Serials and Journals. Retrieved December 18, 2012, from http://www.zentralblatt-math.org/zbmath/journals

López-Illescas, C., De Moya-Anegón, F., & Moed, H. F. (2008). Coverage and citation impact of oncological journals in the Web of Science and Scopus. *Journal of Informetrics*, *2*(4), 304–316. doi:10.1016/j.joi.2008.08.001

IDENTIFICATION METHOD ON LOW QUALITY PATENTS AND APPLICATION IN CHINA

Zhang Mier[1], Hu Suya[2] and Guo Wei[3]

[1]ZHMILL@dlut.edu.cn, [2]HUSUYAY@hotmail.com, [3]GUOWI@hotmail.com
Dalian University of Technology, School of Business Management, 116023 Dalian (China)

Introduction

In the past decade, China has experienced rapid growth in science, technology and economy. It has also made remarkable progress in the field of patents. China on doubt became the fastest growing country. In 2011, the State Intellectual Property Office (SIPO) received 526,412 invention patent applications and China surpassed the U.S. and Japan in patent application for the first time.

Meanwhile, low quality patents are springing up. The negative effects are showing up. Fu (2009) suggests that teachers are accustomed to regard patent filing as means of fulfilling assessment requirements. Zheng et al. (2009) deem that patent system makes applicants to implement the strategy of high quantity and low quality. However, relevant quantitative research is rather rare. It is an urgent task to identify low quality patents.

Identification method of low quality patents

Principles of index selection

When choosing index, we should consider some indicators are influenced by technical features easily. This would affect the applicability of study. Next, the study relies on patent data, but patent documentations vary in different country and database. We should utilize common information. Besides, confronted with big data, we'd better choose some simple and direct indexes. This could enable us to filter out low quality patents precisely.

Based on above analysis, three principles should be satisfied. Firstly universality, the research method should be applied to different fields, different regions and different countries, so as to do in-depth analysis. Secondly calculability, the index should be obtained directly or by calculating. Thirdly feasibility, because of the large scale of patent data, we should choose simple and direct index as much as possible.

Index selection

On the basis of these principles, we should take mutual patent documentation information of main economic entity into consideration. Then citation index could be ruled out. Only U.S. patent documentations contain integrated information of citation. China has no citation information.

Patentees know best about their patent quality. Griliches (1990) claims patentees would choose to maintain patent only when prospective earnings are greater than maintenance fee. Gronqvist (2009) considers the longer maintenance time of a patent, the more worthy it is. Barney (2003) stresses high maintenance rate is in accordance with

high quality. Harhoff et al. (1999) shows that expiration patents are of higher quality than early termination patents. So, the index of maintenance could be used to exclude low quality patents.

The threshold

After index selection, there should be appropriate threshold to work as screening standard. It is the lowest time that identifies low quality patents.

Patent maintenance fee is one of the key factors to determine the threshold. Qiao (2011) holds the view that the ratio of expiration is proportional to patent quality. According to patent laws in China, the amount of patent maintenance fee goes up one stair triennially in the first fifteen years and then only one stair at the last five years.

Tab.1 Patent Maintenance Fee

Period (year)	Annual fee (CNY)	total fee (CNY)
1-3	900	2700
4-6	1200	6300
7-9	2000	12300
10-12	4000	24300
13-15	6000	42300
16-20	8000	82300

In order to determine the threshold, we should consider the upper limit time of invention patents, the average maintenance time of Chinese patents, and the stepped feature of patent maintenance fee. So, we set 6 years as the threshold.

Application example: low quality patents in telecommunication

Research sample

We rely on Chinese patent database and choose invention patents in telecommunication as research sample. According to the IPC code, its class is H04, including 11 subclasses.

Based on the nationality of applicants, Chinese patents could be divided into two categories: domestic patents and foreign patents. It will be conducive to analyze the formation and variation tendency about low quality patents via comparing the domestic and foreign low quality patents.

Tab.2 The Quantity of Invention Patents and Low Quality Patents

Year	Invention patents		Low quality patents(T<6)	
	A	B	A	B
1990	23	143	21	31
1991	26	119	18	19
1992	27	156	20	47
1993	30	87	25	16
1994	16	106	15	25
1995	16	91	14	14
1996	22	92	15	11
1997	23	111	16	20
1998	20	144	10	25
1999	31	248	22	52
2000	149	360	70	65
2001	159	660	66	187
2002	109	1607	37	394
2003	511	3351	129	719
2004	1280	4844	309	1019
2005	1188	3963	248	893

(A=domestic patents, B= foreign patents)

Data processing

We choose PIAS V3.0 as data source. It's a platform instrument combining information retrieval, patent management and analysis. Aiming at Chinese patents, we need to utilize patent information of invention authorization. So, we choose CNIPR database to download data.

Sample data would trace back to 1990 in order to get longer time series and put forward to 2005 because of the threshold. Use the search engine to gather information of the invalid date of patents. Calculate maintenance time of each invalid patent. Figure out the patents which maintenance time less than 6 years.

The results show the number of patents and low quality patents are growing rapidly, especially after 2000. In view of the fast increases in patent authorization, further analysis about the proportion of low quality patents would be necessary.

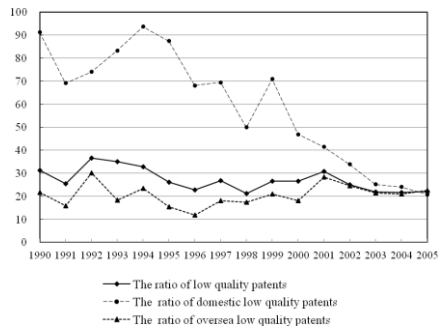


Fig.1 The Trend of Low Quality Patents

Data analysis

With the rapid increase in patent authorization, massive low quality patents are springing up in China. The growing number of low quality patents shows the trend of accelerating. The average annual growth rate in 1990s and 2000s is 10.01% and 53.25% accordingly. Not only are the domestic in vigorous growth, but also the foreign patents.

The ratio of low quality patents fluctuates at 25%. Domestic and foreign ratio of low quality patents approach in 2000s. This illustrates Chinese domestic patent quality is gradually improving. The analysis of patentees shows that leading companies, such as Huawei Technologies Co. Ltd and ZTE Corporation, turning into dominating patentees.

Conclusion

With the rapid growth of patents, patent foam burst in China. Based on the analysis of patent quality and patent maintenance, an identification method and index are proposed. Taking telecommunication patents in China as research samples, this method is employed to identify low quality patents. The results demonstrate that there are large numbers of low quality patents in this field.

Acknowledgments

This research is supported by National Natural Science Foundation of China (71210307037, 71172138) and Natural Science Foundation of Liaoning Province (201202038).

References

Barney, J. A. & Barney, J. R. (2003). Method and System for Rating Patents and other Intangible Assets: US, 6556992.

Fu, Y., Ma, Q. & Sheng, P. Z. (2009). An Analysis on the Status-Quo of Valid Patent in Universi-ties. *Science of Science and Management of Science& Technology*, 8:45-49.

Griliches, Z. (1990). Patent Statistics as Economic Indicators: a Survey. *Journal of Economic Literature*, 28(4):1661-1707.

Gronqvist, C. (2009).The Private Value of Patent Characteristics: Evidence from Finland. *The Journal of Technology Transfer*, 34(2):159-168.

Harhoff, D., Narin, F., Scherer, F. & Vopel, K. (1999). Citation Frequency and the Value of Patented Inventions. *The Review of Economics and Statistics*, 81:511-515.

Qiao, Y. Z. (2011). The Research of the Annual Fee Mechanism of Patent Maintenance. *Studies in Science of Science*, 29(10):1490-1494.

Zheng, Y. P., Dang, X. M. & Yu, l. F. (2009). Study on the Radical Causes in Developing Status of University's Patent Quality Based on the Change of Patent Status in Research Projects. *Science& Technology Progress and Policy*, 26(19):183-186.

IMPACT AND VISIBILITY OF SA'S RESEARCH JOURNALS: ASSESSING THE 2008 EXPANSION IN COVERAGE OF THE THOMSON REUTERS DATABASES

Androniki Pouris[1], Anastassios Pouris[2]

[1]*androniki@hotmail.com*
Faculty of Science, Tshwane University of Technology, Pretoria (South Africa)

[2]*apouris@icon.co.za*
Institute for Technological Innovation, University of Pretoria, Pretoria (South Africa)

Introduction

In South Africa journal evaluation has a long history (POURIS 1986; POURIS AND RICHTER 2000; POURIS 2005).During 2006 the Academy of Science of South Africa produced a new strategic framework for South Africa's research journals (ASSAf 2006). The framework recommends the periodic peer review of the country's journals and a move into an open access system.

During 2008 (Thomson-Reuters 2008) added 700 regional journals in the Web of Science focusing on particular topics of regional interest. The number of South Africa set increased by 19 journals. South Africa with 29 journals in the scientific domain is above New Zealand, Ireland, Mexico and Israel. In the social sciences South Africa is even above countries like Japan and China probably because of language issues. This article aims to answer two questions. First, we aim to compare the performance of the journals indexed by the JCR during 2002 with their performance during 2009 and 2010 and secondly to identify whether the newly added SA journals in the JCR are of similar quality as the pre-existing journals. Both findings are of policy interest and we discuss the consequences of our findings.

Methodology

Two approaches are used for the comparison and assessment of journals – expert opinion (Zhou et al 2002) and citation analysis (Ren and Rousseau 2002).

In order to facilitate comparisons among journals belonging to different scientific disciplines we identify the quartile in which the various journals belong within the disciplines in which the journals are classified in JCR. So if a scientific discipline covers 20 international journals we rank them in a descending order according to their impact factor and we classify them in four quartiles. The top quartile includes the five journals with the highest impact factors.

The Performance of South African Journals

Table 1 shows that out of the 17 journals five journals lost ground in terms of quartiles during the 2002-2009 period. Two of them recovered during 2010.Only one journal improved its performance and moved from the third

quartile to second one – the African Journal of Marine Science. Examination of the impact factors indicates that 12 journals increased their impact factors. However, these increases were insufficient to move them into higher quartiles.

Table 1: Impact factors& Quartiles of Pre-existing SA Journals in JCR-SCI-2002, 2009 & 2010

Journal	IF 2002	IF 2009	Quart 2002	Quart 2009	Quart 2010
AFR ENTOMOL	0.455	0.42	3	4	4
AFR J MAR SCI	0.754	1.52	3	2	3
AFR ZOOL	0.516	0.746	3	3	2
BOTHALIA	0.358	0.242	2	4	3
J S AFR I MIN METAL	0.052	0.216	4	4	4
				3	-
J S AFR VET ASSOC	0.366	0.224	3	4	3
ONDERSTEP J VET	0.506	0.43	3	3	3
OSTRICH	0.149	0.25	4	4	4
S AFR J ANIM SCI	0.381	0.412	3	3	3
S AFR J BOT	0.394	1.08	2	3	4
S AFR J CHEM - S A T	0.265	0.429	4	4	4
S AFR J GEOL	0.659	1.013	4	2	4
S AFR J SCI	0.7	0.506	2	3	2
S AFR J SURG	0.25	0.429	4	4	4
S AFR J WILDL RES	0.224	0.562	4	4	4
SAMJ S AFR MED J	1.019	1.325	2	2	2
WATER SA	0.481	0.911	3	3	3

Table 2: Impact factors & quartiles of new SA journals added in JCR-SCI

Journals	IF 2009	Quartile	Quartile 2010
AFR INVER	1.216	2	3
AFR J HERPET	0.455	4	4
AJAR -A J AIDS	0.569	4	4
INT SPORTM J	0.171	4	4
J S AFR IN C EN	0.125	4	4
QUAEST MATH	0.267	4	4
S AFR J ENO V	0.314	4	3
S AFR J HIV M	0.457	4	4
S AFR J IND E	0.093	4	4
SAJOG - A J OB	0	-	4
SAJP - S A J PS	0.409	4	4
S FORESTS	0.5	4	3

Table 2 shows the new South African journals added in the journal citation reports. With the exception of the *African Invertebrate* which is positioned in the second quartile of the relevant journals with impact factor 1.2, all other journals fall within the fourth quartile of their categories

In order to compare the quality of the pre-existing journals with that of the new journals as it is manifested in their impact factors we undertook a two-sample t-test. The test identified that we cannot reject the null hypothesis that the two sets of journals come from the same population.

Table 3 shows the SA journals in the JCR social sciences citation index. The * indicate the journals which were indexed during 2002. South Africa is represented by 16 journals during 2009 (4 during 2002). A two-sample t-test in the two sets of journals again indicates that all journals are coming from the same population as far as their impact factors are concerned.

Table 3: Impact factors and quartiles of SA journals in JCR-SSCI

Journal Name	Impact Factor 2009	Quart 2009	Quart 2010
AJAR - AFR J AIDS RES	0.569	4	4
ANTHROPOL S AFR	0.222	4	4
EDUC CHANGE	0.17	4	4
LEXIKOS	0.667	3	2
PERSPECT EDUC*	0.387	3	4
POLITIKON - UK	0.216	4	4
S AFR GEOGR J	0.207	4	4
S AFR J BUS MANAG	0.146	4	4
S AFR J ECON*	0.248	4	4
S AFR ECON MANAG S	0.082	4	4
S AFR J EDUC	0.308	4	3
S AFR J HUM RIGHTS	0.692	3	-
S AFR J PSYCHOL*	0.347	4	4
SAJP - S AFR J PSYCHI	0.409	4	4
SO AFR LINGUIST APPL	0.066	4	4
SOC DYNAMICS*	0.237	3	4

References

ASSAf (2006) Report on a strategic approach to research publishing in South Africa, Pretoria

POURIS A. (1986). The South African Journal of Science: A Bibliometric Evaluation" *South African Journal of Science* 82(August): 401-2

POURIS A and RICHTER L. (2000) Investigation into state funded research journals in South Africa, *South African Journal of Science* 96 (March) 98-104

POURIS A. (2005). An assessment of the impact and visibility of South African journals, *Scientometrics* 62 (2): 213-222

REN S, ROUSSEAU R, International visibility of Chinese scientific journals, *Scientometrics*, 53 (3) (2002) 389–405.

THOMSON REUTERS PRESS RELEASE (2008) http://science.thomsonreuters.com/press/2008/

D. ZHOU, J. MA, E. TURBAN, N. BOLLOJU, A fuzzy set approach to the evaluation of journal grades, *Fuzzy Sets and Systems*, 131 (2002): 63–74

IMPACT OF BRAIN DRAIN ON SCIENCE PRODUCTION: A CASE STUDY OF IRANIAN EDUCATED MIGRANTS IN THE CONTEXT OF SCIENCE PRODUCTION IN CANADA

Kayvan Kousha[1] , Ashraf Maleki, Mahdieh Hatami, Mahsa Ganji, Sheida Vanoiee, Hamideh Asadi, Razieh Mehdizadeh-Maraghi, Alireza Badrloo, Samira Goodarzi, Shadi Moshtagh, Parisa Sepehr-Ara, Nima Gholami, Akram Siyahi, Mohsen Tavakoli[2]

[1] *k.kousha@wlv.ac.uk*
Statistical Cybermetrics Research Group, School of Technology, University of Wolverhampton, Wulfruna Street, Wolverhampton WV1 1LY (UK) and Scientometrics Department, University of Tehran (Iran)

[2] *malekiashraf@ut.ac.ir*
Master Students of Scientometrics, Library and Information Science Department, University of Tehran, Enghelab sq., Tehran (Iran)

Introduction

Emigration of highly educated people from developing to developed countries has surged during the last decade (Lowell and Findlay, 2001). Canada is also one of the countries that permits a large amount of immigrations from different nations. According to the *Citizenship and Immigration of Canada* (CIC), Iran has been one of the top 10 countries in terms of immigration rate over the past decade and about 64,570 Iranians, mostly highly skilled, have been granted permanent residence of Canada during 2002-2011 (CIC, 2012). Other statistics of CIC also show that total entries of Iranian students to Canada have increased from 445 students in 2002 to 1,252 in 2011. Several investigations have studied reasons for Iranians' migration to developed countries (e.g., Garousi, 2003; Entezarkheir, 2005). Nevertheless, it is not fully known how this may influence scientific productivity of destination countries.

Reviewing scientific publications of Canada in engineering, we noticed many Persian authors' names with Canadian affiliations. Having said that a huge number of highly educated Iranians have migrated to Canada during the past decade (CIC, 2012), the objective of this research is to determine the share of Iranian authors' contribution in scientific publications of Canada. As a case study, we limited our study to scientific publications of Canada in engineering (2005-2011). We also manually checked a sample of Web CVs for Canadian-affiliated authors (corresponding authors, see below) with Persian names to assess their educational or occupational backgrounds.

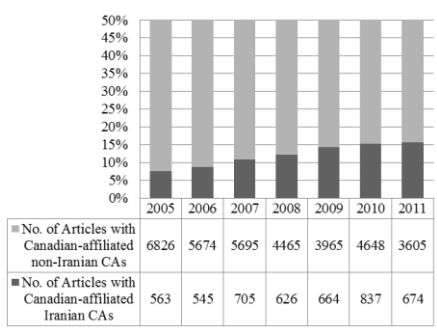
Methodology

We used Elsevier's Scopus database to extract engineering articles published in 2005-2011 by Canada. Ultimately, 39,493 articles with Canadian

corresponding authors were collected for the purpose of the study.

Identifying Persian names

To assess Iranians' contribution in scientific publications of Canada, 13 master students went through an extensive human labour task and manually checked 39,493 articles for corresponding authors with Persian names. We used corresponding authors (CA) as a better indicator for the purpose of study. Generally Iranian names have some special characteristics and could be easily identified by native Persian people. For instance, suffixes like -pour, -ni(y)a, -zadeh, -far, -nejad in Iranian last names helped us to identify Iranian authors. However, as the authors' last names were not always enough to recognize Iranian names, we also checked the first names of the authors either as recorded in Scopus outputs or through online CVs, if necessary. Note that we visited web CVs of a sample of 320 (16.7%, $\alpha = 0.05$) authors to determine the accuracy of method, finding that there is only 1.6% error in identification of Persian authors' names, which proves it is reliable.

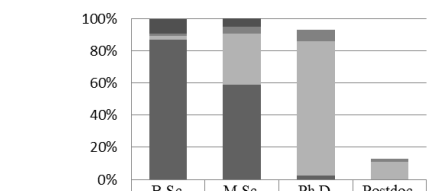


	2005	2006	2007	2008	2009	2010	2011
No. of Articles with Canadian-affiliated non-Iranian CAs	6826	5674	5695	4465	3965	4648	3605
No. of Articles with Canadian-affiliated Iranian CAs	563	545	705	626	664	837	674

Figure 1. Number(%) of Scopus publications with Persian names and Canadian affiliation in engineering (2005-2011)

Results

Figure 1 and 2 summarise main results of the study. The overall result indicates that almost 1,908 unique corresponding authors with Iranian names have contributed to 4,614 scientific articles of Canada in engineering during 2005-2011. Most importantly, according to Figure 1 we can estimate that the proportion of articles with Canadian-affiliated Iranian CA to all the articles with Canadian CA is 12% (4,614 articles). It also suggests a constant increase in the proportion of publications with Iranian corresponding authors and Canadian affiliation from 8% in 2005 to 16% in 2011.



	B.Sc.	M.Sc.	Ph.D.	Postdoc.
Unknown	9%	5%		
Other countries	2%	4%	7%	2%
Canada	2%	32%	83%	11%
Iran	87%	59%	2%	0%

Figure 2. Canadian-affiliated Iranian authors' educational destination country in various levels.- Unknown: unavailability of information.

Figure 2 gives more details about the educational backgrounds of corresponding authors with Persian names based on manual checking of 315 sampled online CVs. It indicates that about 87% (274 authors) of the sampled authors have received their BS degrees in the Iranian universities, whereas this is only 4% (12 authors) for other non-Iranian countries (including Canada with 2%). Note that 9% of CVs did not include any information about educational backgrounds of authors. This confirms that many Iranian highly educated human capitals have migrated

to Canada after undergraduate education and could substantially have a significant impact on Canadian research productivity. Figure 2 also reports that about 83% of authors have received their PhD in the Canadian universities.

Further analysis of CVs revealed that 67% and about 50% Iranian authors received their BS and MS degrees respectively from top Iranian universities in engineering such as Sharif University of Technology, University of Tehran, Isfahan University of Technology, Amirkabir University of Technology, and Iran University of Science and Technology.

As shown in Table 1, content analysis of sampled CVs disclosed that 76% of authors have occupations at organizations or universities abroad (mostly in Canada) and only 6% mentioned positions in Iran (return of migrates to the native land). This suggests huge brain drain rate rather than other patterns of scientific communication such as mobility, or brain exchange.

Table 1. Current occupational position of tested sample

Current occupational position	No. (%) of tested sample
Occupied at organizations or universities abroad	240 (76%)
Students at universities abroad	48 (15%)
Occupied at organizations or universities in Iran	18 (6%)
Unknown	9 (3%)
Total	315 (100%)

Conclusions

Although, scientific migration has benefits for science and research communication, an extensive outgoing of skilled migrants and human capitals from developing to developed countries may have negative and sometimes damaging impact on exporting nations. The method here was useful to assess how migration of Iranian scholars to other countries can play a major role in scientific production of hosting country and could be helpful for the future research policy decision making in both sides. Bibliometric methods can help to explore who is winner and loser and what are the scientific and economic consequences for developing countries in losing highly skilled human capital assets.

References

CIC (2012). *Canada facts and figures*. Ontario. Retrieved from www.cic.gc.ca/english/resources/statistics/menu-fact.asp

Entezarkheir, M. (2005). Why is Iran experiencing migration and brain drain to Canada? *34th annual conference of the Atlantic Canada Economic Association, papers & proceedings* (pp. 1–30).

Garousi,V. (2003). A survey on the immigration of Iranian experts and the elite: reasons, losses and possible solutions. *Scientific seminar on the discourse of overseas Iranian youth*. Tehran, Iran.

Lowell, BL and Findlay, AM. (2001). Migration of highly skilled persons from developing countries: impact and policy responses. Geneva: International Labour Office, www.ilo.org/public/english/protection/migrant/download/skmig-sr.pdf

AN INDEX TO QUALIFY HUMAN RESOURCES OF AN ENTERPRISES CLUSTER

Yu Liu[1] and Kun Ding[1]

[1] *rachelliuyu@hotmail.com, dingk@dlut.edu.cn*
WISElab, Dalian University of Technology, No.2 Linggong Road, Ganjingzi District,116024, Dalian (China)

Introduction

Since it was firstly and officially proposed by Hirsch (2005), the h-index and then the g-index by Egghe (2006) are gaining continuous focus and constantly studies both theoretical and empirical. The h-index can simultaneously represent the qualitative and quantitative meanings for scientific indicators such as patents, authors, publications, journals. Studies on h-index have seen some theoretical advances (Egghe & Rousseau, 2006; Glanzel, 2006; Hirsch, 2010), but they are still confined in qualifying the S&T impact or outputs of researchers (Oppenheim, 2007), journals (Braun, Glanzel & Schubert, 2006), institutions (Molinari & Molinari, 2008) or groups (Van Raan, 2006). Seldom study or extension of h-index was developed in other aspects or subjects.

Inspirited by the mechanism of h-index, we presented l-index for the first time to demonstrate human resources distribution (HRD) in a set of enterprises, and initially defined 11 l-indexes corresponding to 11 types of staffs. The set of enterprises are sharing one or more identities, for instance, the major industries. We aim to describe and evaluate the integrative technological innovative capacities of an enterprise cluster from human resources perspective, to describe the innovative resources of enterprises using a more visual, simple but meaningful alternative rather than using mass data. Inheriting the identities of h-index that both qualitative and quantitative, l-index is not so obscure to be understood. The conception of l-index combines the bibliometrics' thinking and approaches with the studies on enterprises, making bibliometrics' thinking be of more practical significance in enterprises research and contributing to the development of cross-disciplines.

Theoretical and practical definitions

Definition

In a set of enterprises sharing one or more identities, if at least x% ones of all meet a condition that each of them owns a certain type of staffs by x% or larger, the l-index for this type of staffs of the enterprises' set equal to x.

11 l-indexes

We initially developed 11 l-indexes, each of them corresponded to a type of staffs: foreign experts, researchers, managers, other function staffs, staffs with PhD degree, staffs with master degree, staffs with bachelor degree, staffs with degree below bachelor, staffs with senior professional title, staffs with medium-degree professional title, and staffs with junior professional title. We sequentially remarked them by T1 to T11. All the 11 l-indexes -l_1 to l_{11}-could compose an integral description for the

staffs' distribution of the enterprises cluster.

For example, if there are no less than 60 out of 100 software R&D enterprises, the ratio of researchers for all the 60 ones is 60% or larger, then the value of the researcher l-index of these 100 software R&D enterprises is equal to 60, i.e. l_2 of these 100 software R&D enterprises is equal to 60. It is notable that the l-indexes are explored and exploited to evaluating the overall HRD in a set of enterprises rather than in the individual.

Feasibility analysis

3 variables

We defined the general proportion, the average proportion, and the l-index as Var1, Var2 and Var3 respectively, all of them are referring to a certain type of staffs distributing in a set of enterprises with varied populations. Var1 represents the proportion that all the staffs of the type account for of whole population, Var2 stands for mean value of the proportions that the staffs of the type account for of every population, Var3 represents l-index for the staffs of the type. Variables' values of T1-T11 (values of l_1-l_{11}) of 23 clusters (1 representing all the 256 Dalian enterprises[183], and 7/6/9 respectively representing attributions/ certifications/ major industries) were calculated and illustrated in Fig.1- Fig.3.

Methods

To testify the validities and accuracies of the l-indexes, methods were adopted, e.g. correlation analysis and variances comparisons.

Processes and results

To begin with, we analyzed relations between the 3 variables and found out that each one was significantly related to another with values all close to 1, and Var3 was more related to Var2, proving that (1) l-indexes can surely replace traditional evaluating indicators, and that (2) Var3 can be represented by Var2 better than by Var1, implying Var3 can precisely reflect the average level of HRD rather than roughly represent an approximate value. Then we compared variances of the 3 variables and every couples. It turned out that the variance of Var2-3 couple was generally smaller than that of Var1-3 couple for most staffs' types, assistant proving l-indexes' credibility. Meanwhile we have abstracted some interesting conclusion from the outcomes that R and BAC were the most significant human resources components of all clusters, and that BBAC and staffs with a professional title were not generally distributing in these enterprises.

Empirical studies

When we testified the feasibilities of l-indexes, we made all the values of 11 l-indexes worked out as well (Fig.1-Fig.3). Not only we made an overview of HRD by l-indexes, we also testified 5 hypothesizes about HRD regulations that (1) T2 are much more than any other while T3 are to the contrary; (2) T7 occupy the major status accompanying with smaller but stable T5 and T6 proportions; (3) T1, T2 and T3 are generally better educated than T4; (4) the grades of professional titles are positively related to the educational background; (5) staffs with professional titles are not commonly distributing in any enterprises type. They were once proved by Var1 and Var2 and testified again by l-indexes. At last, it was proved that T1-T11 composed an maximum

[183] Year of data collecting: 2012.

independent group by each of the 3 variables.

Conclusion and prospection

Theoretical principles and empirical studies of l-indexes were worked out. Though more testing and moderating would still be needed, this series of l-indexes for human and the hiding thinking would absolutely benefit our further study. We're hoping and working to develop other indexes following the mechanism of h-index such as leaders in enterprises.

Acknowledgments

There are loads of people I want to give thanks to. Thanks Prof. Ding for keeping giving tutorial to me, thanks my family for they always mentally support me, and thanks Doc. Gao as well as other colleagues of mine for constructively advising and inspiring me.

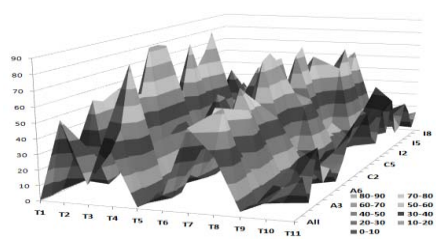


Figure 1. Var1's values of T1-T11 of 23 clusters

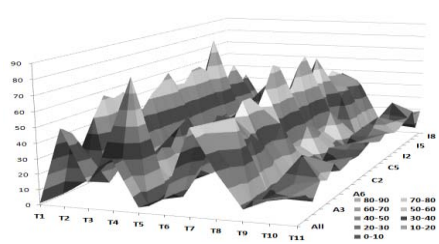


Figure 2. Var2's values of T1-T11 of 23 clusters

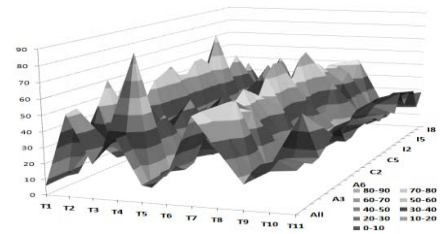


Figure 3. Values of l_1-l_{11} of 23 clusters

References

Braun, T., Glanzel, W., & Schubert A. (2006). A Hirsch-type Index for Journals. *Scientometrics,* 69(1), 169-173.

Egghe, L. & Rousseau, R. (2006). An informetric model for the Hirsch-index. *Scientometrics,* 69(1), 121-129.

Egghe, L. (2006). Theory and practice of the g-index. *Scientometrics*, 69(1), 131-152.

Glanzel, W.(2006). On the h-index - a mathematical approach to a new measure of publication activity and citation impact. *Scientometics,* 67(2), 315-321.

Hirsch, J.E. (2005). An index to quantify an individual's scientific research output. *Proceedings of the National Academy of Sciences,* 102(46), 16569-16572.

Hirsch, J. E. (2010). An index to quantify an individual's scientific research output that takes into account the effect of multiple co-authorship. *Scientometrics,* 85, 741-754.

Molinari, J. F. & Molinari, A. (2008). A New Methodology for Ranking Scientific Institutions. *Scientometrics,* 75(1), 163-174.

Oppenheim, C. (2007). Using the h-Index to rank influential British researchers in information science and librarianship. *Journal of the American Society for Information Science and Technology,* 58(2), 297-301.

Van Raan, AFJ. (2006). Comparison of the Hirsch-index with standard bibliometric indicators and with peer judgment for 147 chemistry research groups. *Scientometrics,* 67(3), 491-502.

AN INTERPRETABLE AXIOMATIZATION OF THE HIRSCH-INDEX

Denis Bouyssou[1] and Thierry Marchant[2]

[1] *bouyssou@lamsade.dauphine.fr*
Université Paris Dauphine - CNRS, Place du Maréchal de Lattre de Tassigny, F-75775 Paris Cedex 16 (France)

[2] *thierry.marchant@ugent.be*
Ghent University, H. Dunantlaan 1, 9000 Ghent (Belgium)

Introduction

In the last ten years, many bibliometric indices have been proposed for comparing and/or evaluating scientists. Among theses indices, the Hirsch-index (or h-index) is probably the most popular one. Since it is not clear which index is best, some researchers have tried to enrich the debate by analyzing various indices from an axiomatic perspective. This stream of research has delivered six[184] axiomatizations of the h-index: Woeginger (2008a,b), Quesada (2009), Quesada (2010), Quesada (2011), Miroiu (2013). They pave the way towards a better understanding of the h-index, but they are not completely satisfactory. That is why we propose a new axiomatization.

Existing axiomatizations and their shortcomings

Consider an index h' defined as 100 times the h-index. Is it worse or better than the h-index? This question is obviously irrelevant, just like asking whether measuring distances in meter is better than in centimeters. Unfortunately, all aforementioned papers axiomatize the h-index instead of considering the family of all indices h' such that h' is equal to k times h. The axioms in these papers are therefore stronger then needed: they implicitely, or sometimes explicitely, state that the h-index of a scientist with one publication and one citation is one, while this actually does not matter.

We now discuss some specific problems.

Woeginger (2008a)

Theorem 4.1 in Woeginger (2008) characterizes the h-index by three axioms called A1, B and D. Axiom A1 is stated as follows: "If the $(n+1)$-dimensional vector y results from the n-dimensional vector x by adding a new article with $f(x)$ citations, then $f(y) \leq f(x)$." Although this axiom is mathematically fine, we claim that it is not interpretable. Indeed, an axiom is a condition imposed on the index f, where f is any index, not necessarily the h-index. So, when we read this condition, we may not suppose that f is the h-index. It could be the square of the number of papers or the logarithm of the total number of citations, … It does therefore

[184] Notice that Marchant (2009) does not belong to this list because it does not axiomatize the h-index, but the ranking induced by the h-index. Burgos (2010) and Gagolewski (2011) do also not belong to the list because they do not axiomatize the h-index but a family of indices containing the h-index.

not make sense to say "if we add a new paper with $f(x)$ citations, then …" Why would we find such a condition (normatively) appealing if we do not know what $f(x)$ represents? Axiom D has the same problem.

Woeginger (2008b)

This paper assumes that a bibliometric index must be a non-negative integer. This is very restrictive and difficult to motivate. It also uses axiom A1 as in Woeginger (2008a).

Quesada (2009)

Here, Axiom A1 imposes that $f(x)$ lies between (a) the minimum of the number of cited papers and the smallest number of citations (not taking uncited papers into account), and (b) the minimum of the number of papers and the largest number of papers. This is a complex condition. Actually, it combines several conditions.

Miroiu (2013)

This paper also assumes that a bibliometric index must be a non-negative integer. Besides, it uses some axioms (CPI, PR, CCI and CJ) that suffer the same problem as axiom A1 in Woeginger (2008a): they compare an unspecified index to a number of citations. This is not interpretable as long as we do not know which index is considered.

A new axiomatization

Among the aforementioned axiomatizations, those of Quesada seem most promising. We propose hereunder a list of axioms, inspired from those of Quesada, and we use them to axiomatize the family of all indices h' such that h' is equal to k times h.

Non-Triviality: there are scientists x, y such that $f(x) \neq f(y)$.

Zero: scientists with no paper or only uncited papers have an index equal to 0.
Tail Independence: suppose x and y have the same number of papers and $f(x) = f(y)$. Suppose both publish an additional paper, with the same number of citations, at most equal to the number of citations of the least cited paper of x and y. Then $f(x') = f(y')$.
Square Upwards: suppose x has m papers, each with m citations. Suppose x gets some additional citations. Then $f(x') = f(x)$.

Square Rightwards: suppose x has m papers, each with m citations. Suppose x publishes some additional papers with at most m citations. Then $f(x') = f(x)$.
Homothety: suppose x has m papers, each with m citations, and y has one paper, with one citation. Then $f(x) = m f(y)$.

Theorem : an index f satisfies Non-Triviality, Zero, A2 (Quesada), Tail Independence, Square Upwards, Square Rightwards and Homothety iff f is the h-index multiplied by some positive real number.

Compared to Theorem 3.1 in Woeginger (2009), our Theorem is more interesting because it axiomatizes the family of all h-indices. Moreover, it uses simpler axioms. For instance, A1 has been splitted into Square Upwards and Square Rightwards.

References

Burgos, A. (2010). Ranking scientists. UMUFAE Economics Working Papers, Universidad de Murcia, 2.
Gagolewski, M. (2011). Possibilistic analysis of arity-monotonic aggregation operators and its relation

to bibliometric impact assessment of individuals. *International Journal of Approximate Reasoning*, 52, 1312-1324.
Marchant, T. (2009). An axiomatic characterization of the ranking based on the h-index and some other bibliometric rankings of authors. *Scientometrics*, 80, 325-342.
Miroiu, A. (2013). Axiomatizing the Hirsch index: Quantity and quality disjoined. *Journal of Informetrics*, 7, 10-15.
Quesada, A. (2009). Monotonicity and the Hirsch-index. *Journal of Informetrics*, 3, 158-160.
Quesada, A. (2010). Further characterizations of the Hirsch index. *Scientometrics*, 87, 107-114.
Quesada, A. (2011). More axiomatics for the Hirsch index. *Scientometrics*, 82, 413-418.
Woeginger, G. H. (2008a). An axiomatic characterization for the Hirsch-index. *Mathematical Social Sciences*, 56, 224-232.
Woeginger, G. H. (2008b). A symmetry axiom for scientific impact indices. *Mathematical Social Sciences*, 2, 298-303.

INTERPRETING EPISTEMIC AND SOCIAL CULTURAL IDENTITIES OF DISCIPLINES WITH MACHINE LEARNING MODELS OF METADISCOURSE

Bradford Demarest [1] and Cassidy R. Sugimoto [2]

[1] *bdemares@indiana.edu*

[2] *sugimoto@indiana.edu*

Indiana University, School of Library and Information Science, 1320 East 10th Street, LI 011; Bloomington, IN 47405-3907 (United States)

Introduction

Scholars of academic disciplinary differences note that these differences encompass not only differences in topic, but also in "sanctioned social behaviours, epistemic beliefs, and institutional structures of academic communities" (Hyland, 2004, p.2). Discerning these disciplinary differences through text analysis has been mostly limited to qualitative or corpus methods, and has mostly excluded machine learning based methods (except Argamon, Dodick, and Chase [2008]). Further, the research of disciplinary beliefs and behaviors as reflected in discourse has almost entirely focused on research articles, despite genre scholars' contention that writing patterns can vary widely among genres (Hyland, 2004). The current study seeks to address both of these gaps in the literature by modeling social and epistemic differences through a machine-learning approach among philosophy, psychology, and physics based on metadiscourse (words or phrases in a text that position the author, the text itself, and the reader in relation to one another [Hyland, 2005]) used in dissertation abstracts. A previous study of dissertation abstracts (Demarest & Sugimoto, 2013) explored the metadiscourse use of disciplines in contrast to multiple other disciplines, but left aside contrasts between specific disciplines, which the current study undertakes. The findings of both studies support previous qualitative and corpus-based studies (Becher, 1987; Hyland, 2008), which established epistemological and social differences among hard sciences, social sciences, and humanities. The current study asks two research questions: 1) Can metadiscourse be used via a machine learning approach to model disciplinary differences in the dissertation abstract genre in ways that can empirically test existing theories of disciplinary culture? 2) How could such a model offer a way to position disciplines in relation to one another in a network of metadiscursive (and thus social and epistemic) proximity?

Methods

Dissertation abstracts from 1990-2008 in a subject category containing one of the strings "physics", "psychology", or "philosophy" were taken from the ProQuest database (excluding abstracts containing two or more identifying

strings). The data set was then balanced for discipline frequencies: the discipline with the lowest number of abstracts was found (philosophy) and abstracts were randomly sampled from the other two disciplines at a frequency based on the proportion of each larger discipline's abstract count to philosophy's, yielding 49746 abstracts (16602 from philosophy, 16592 from physics, and 16552 from psychology).

For the set of features, 316 words and phrases expressing interaction from Hyland (2005) were collected; removing 13 cross-category duplicates and adding four Americanized spellings for terms ("analyse", "realize", "realizes", "realized") yielded 307 features. After collecting relative frequencies for each feature for each abstract in the sample, the WEKA machine-learning program (Hall et al., 2009) version 3.6.6 was used to create a multi-class sequential minimal optimization (SMO) support vector model (Platt, 1998), which combines models of the three pairwise combinations of disciplines. The PolyKernel kernel was used with default settings and, based on results from the CVParameterSelection optimization algorithm, a complexity parameter value of 10. The resulting model was then tested using ten-fold cross-validation for classification accuracy.

Table 1. Accuracy rates identifying disciplines across pairwise models, and overall average.

Discipline	*Accuracy (%)*
Philosophy	81.9
Psychology	77.2
Physics	80.5
Average	***79.9***

Results

Table 1 presents the accuracy rate by percentage for each discipline, as well as the accuracy percentage averaged across all three disciplines.

Table 2 contains a confusion matrix presenting counts of correct and incorrect classifications of each discipline. The leftmost column designates actual disciplines of abstracts, with other columns containing the numbers of abstracts classified as the column's discipline.

Table 2. Confusion matrix of disciplines.

Discipline	*Philosophy*	*Psychology*	*Physics*
Philosophy	13289	1395	1918
Psychology	1696	12242	2614
Physics	853	1519	14220

Features with weights having absolute values greater than 10 are discussed below as indicators of disciplinary differences..

Discussion and Conclusions

The first research question asks, "Can metadiscourse be used via a machine learning approach to model disciplinary differences in the dissertation abstract genre in ways that can empirically test existing theories of disciplinary culture?"

The psychology-philosophy model found features indicating philosophy, such as "argue", "claim", and "my", suggest philosophy to be per Becher (1987) interpretive and humanistic. Terms such as "measure", "indicate", and "observe" most strongly indicate psychology, suggesting that it is comparatively quantitative, per Becher's (1987) description of pure sciences.

Feature weights in the physics-philosophy model also indicate that philosophy is person-oriented ("my", "the author", and "one's") subjective and interpretive ("think", "claim", "know", "feel"), while physics is quantitative, explanatory, and empirical ("observe", "measure", "sure").

The physics-psychology model feature weights suggest physics to be relatively quantitative ("calculate"), cumulative ("known"), and empiricist orientation ("observe", "show"), while psychology is relatively subjective (e.g. "assess", "claim", "you", and "mine"), or studying a subjective phenomenon ("feel").

Regarding the second research question ("How could such a model offer a way to position disciplines in relation to one another in a network of metadiscursive [and thus social and epistemic cultural] proximity?"), Table 3 provides proximity measures between disciplines, derived from Table 2 by translating counts into normalized inverse indicators of distance. Table 3 reveals the greatest distance from physics to philosophy, and the shortest distance from psychology to physics. These distances suggest the relative ability of each discipline to incorporate the epistemic and social terminology of other disciplines, either instrumentally (e.g., in a quantitatively inclined wing of psychology) or topically (e.g., in the philosophy of science).

Table 3. Normalized inverse values of confusion matrix

Discipline	*Philosophy*	*Psychology*	*Physics*
Philosophy	1.25	11.90	8.66
Psychology	9.76	1.35	6.33
Physics	19.45	10.92	1.17

Future Research

The scientometric community has historically analyzed differences and similarities among disciplines, but at the disciplinary level has relied on the analysis of concepts, producers, and artefacts to expose the intra- and inter-disciplinary terrain (Borgman, 1989). The study of metadiscourse patterns adds a new object of analysis to this list, and can be used to investigate questions about the basis, structure, and evolution of scientific communities.

Acknowledgements

We would like to thank the National Science Foundation (Grant No. SMA-1208804) for financial support, as well as ProQuest for sharing dissertation data with us for research purposes.

References

Argamon, S., Dodick, J., & Chase, P. (2008). Language use reflects scientific methodology: A corpus-based study of peer-reviewed journal articles. *Scientometrics*, 75(2), 203–238. doi:10.1007/s11192-007-1768-y

Becher, T. (1987). Disciplinary discourse. *Studies in Higher Education*, 12(3), 261–274.

Borgman, C. L. (1989). Bibliometrics and Scholarly Communication Editor's Introduction. *Communication Research*, 16(5), 583–599. doi:10.1177/009365089016005002

Demarest, B. & Sugimoto, C.R. (2013). Using machine learning models to interpret disciplinary styles of metadiscourse in dissertation abstracts. Poster presented at *iConference 2013: Scholarship in Action*, Forth Worth, TX.

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., & Witten, I.H. (2009). The WEKA data mining software: An update. *SIGKDD Explorations* 11(1), 10-18.

Hyland, K. (2004). *Disciplinary Discourses: Social Interactions in Academic Writing*. Ann Arbor, MI: University of Michigan Press/ELT.

Hyland, K. (2005). Metadiscourse: Exploring Interaction in Writing. New York, NY: Continuum.

Hyland, K. (2008). Disciplinary voices: Interactions in research writing.

English Text Construction, 1(1), 5–22. doi:10.1075/etc.1.1.03hyl

Platt, J. (1998). Fast training of support vector machines using sequential minimal optimization. In B. Schoelkopf, C. Burges, and A. Smola (Eds.), *Advances in Kernel Methods - Support Vector Learning*. Cambridge, MA: MIT Press.

AN INVESTIGATION OF SCIENTIFIC COLLABORATION BETWEEN IRAN AND OTHER MENA COUNTRIES AND ITS RELATIONSHIP WITH ECONOMIC INDICATORS

Maryam Fayazi[1]

[1] *mfayazi@alumni.ut.ac.ir*
University of Tehran, Faculty of library and Information Science, Enghelab St., Tehran (Iran)

Introduction

It seems that the effect of collaboration on citation rate maybe depend on the cooperating country (Kim, 2001;) or on the discipline (Frederiksen, 2004). Scholars have differing opinions about the factors can drive researchers to collaborate. (Wagner, Brahmakulam, Jackson, Wong & Yoda, 2001) found that geographic proximity, history, common language, economic factors, research equipment, databases, and laboratories are factors influencing international collaboration. On the one hand Price (1963, p. 168) claimed that collaborative authorship reflects more economic than intellectual dependence. Others have argued that co-authorship reflects mutual intellectual and social influence (Edge, 1979; Stokes & Hartley, 1989). Kim's (2005) research has shown that the expenditure on R&D correlates negatively with internationally collaborative publications in Korea Acosta, Coronado, Ferrandiz & Leon (2011) found that differences in scientific resources between regions are relevant in explaining academic scientific collaborations. Iran has political and economic interaction and common culture and religion with Middle East and North Africa (MENA) countries.

Objectives

The objectives of this study are: (a) to investigate the extent of scientific collaboration between researchers from Iran and other MENA countries as reflected in the WOS database during 1999-2009; (b) to analyse the relationship between the number of authors and the number of times an article is cited for the period of study; and (c) to examine the relationship between (i) the per capita GDP and (ii) the GERD/GDP ratio of each country and its co-authorship with Iran.

Methods

The definition of the MENA region follows the World Bank delineation which includes: Algeria, Bahrain, Egypt, Iran, Iraq, Jordan, Kuwait, Lebanon, Libya, Morocco, Oman, Qatar, Saudi Arabia, Syria, Tunisia, Turkey, United Arab Emirates (UAE) and Gaza, & Yemen (World bank, 2011). We added to this definition Turkey and removed it Djibouti, Israel and West Bank. The scientific production, co-authored publications, and citation, for 1999-2009, were retrieved from the WOS database (SCI, SSCI, Conference Proceedings CI–Science, Conference Proceedings CI-Social Science and Humanities). Scientific productions

were retrieved by the formula CU=* in advanced search. The star can be each of the MENA countries. Iranian co-authored publications with each of the MENA countries was retrieved by the formula CU=Iran AND CU=* and then we used Boolean OR to combine the search results. We used Subject category of Essential Science Indicator. The Gross domestic product (GDP) based on purchasing-power-parity per capita (units Current international dollar) of each countries were retrieved from World Economic Outlook Databases (WEO, September 2011) of International Monetary Fund for 1999-2009. The Gross domestic expenditure on research and development (GERD)/GDP ratio retrieved from UNESCO Institute for Statistics for 1999-2009. All data were retrieved in November 2011.

Table 1. MENA countries scientific productions and rate of its collaboration with Iran during 1999-2009

Country	Scientific production	co-authored publications
Turkey	175689	299 (0.17)
Kuwait	7739	50 (0.646)
UAE	8279	49 (0.592)
Lebanon	7174	41 (0.572)
Egypt*	43425	39 a(0.09)
Saudi Arabia	21852	32 (0.146)
Syria	2003	29 (1.448)
Jordan	8846	22 (0.249)
Qatar	1769	22 (1.244)
Morocco	14141	21 (0.149)
Algeria	11864	19 (0.16)
Oman	3821	19 (0.497)
Tunisia	18000	16 (0.089)
Iraq	1683	11 (0.654
Bahrain	1277	6 (0.470)
Libya	1161	6 (0.517)
Yemen	659	2 (0.303)
Gaza/Palestine	0	0

* Iran as a membership of some international organization has scientific collaboration with other countries, e.g. Iran has scientific collaboration with Egypt through ICARDA.

Results and discussion

There were 637 co-authored publications in WOS for 1999-2009 (see table 1).

Analysis of data indicated that there is a significant and moderate (N=18, Sing.=0.004, Level=0.01, rho=0.637**) relationship between scientific productions of each MENA countries and the rate of its collaboration with Iran. It suggests that Iranian researchers are more eager to collaborate with countries that have high scientific potential.

The nonparametric tests revealed a significant and weak (N=637, Sing.=0.000, Level=0.01, rho=0.213 **) relationship between the number of authors and the numbers of times an co-authored paper is cited in other papers.

Table 2. Correlation test between GDP per capita of each country and its co-authorship with Iran

Country	Correlation coefficient	Sign.	Correlation r value
Lebanon	Pearson	0.000	0.951**
Turkey	Spearman	0.000	0.909**
Oman	Spearman	0.000	0.908**
Syria	Spearman	0.000	0.871**
Egypt	Spearman	0.001	0.869**
Kuwait	Pearson	0.001	0.863**
UAE	Spearman	0.002	0.823**
Bahrain	Spearman	0.004	0.787**
Algeria	Pearson	0.005	0.777**
Morocco	Spearman	0.007	0.761**
Tunisia	Spearman	0.01	0.732*
Jordan	Spearman	0.013	0.719*
SaudiArabia	Pearson	0.13	0.715*
Libya	Spearman	0.085	0.543
Qatar	Spearman	0.596	0.18

Most of co-authored publications were in the fields of physics (123 records), clinical medicine (98), chemistry (90). As Kim (2005) mentioned in his research, fields of physics, chemistry, and clinical medicine need very costly

research laboratories that are exceeding the budgeting of single countries.

The relationship between GDP and the GERD/GDP ratio of each country and its co-authorship with Iran, have been investigated for countries which their GDP and GERD/GDP ratio data were available. If the frequency distribution of data was normal, we used Pearson otherwise we used Spearman. The ** and * show that there is a correlation in the level of (0/01) and (0/05) (see table 2) and (see table 3).

Table 3. Correlation test between GERD/GDP ratio of each country and its co-authorship with Iran

Country	Correlation coefficient	Sign.	Coefficient r value
Turkey	Spearman	0.002	0.827**
Tunisia	Spearman	0.01	0.732**
Kuwait	Pearson	0.013	-0.719*
Egypt	Spearman	0.408	0.342
SaudiArabia	Spearman	0.52	0.342

Conclusion

With respect to the result of this study, it seems that the cultural common interests play few roles in creation of co-authored publications and economic factors and political relationship between countries are more effective in it. In MENA area Turkey and Saudi Arabia have a high scientific potential and scientific collaboration with them may be useful for Iran. The results of the study showed that although Iran and Turkey are neighbour countries, the rate of their scientific cooperation was very low. Following the result of this research, scientific policymakers should invest in the top scientific fields and common scientific productions of each of MENA countries in the future, special with countries that have high scientific potential and have strong positive relationship between the amount of their scientific collaboration with Iran and their GPD or GEDR/GDP ratio.

References

Acosta, M., Coronado, D., Ferrandiz, E, & Leon, M. (2011). Factors affecting inter-regional academic scientific collaboration within Europe: The role of economic distance. *Scientometrics,* 78(1), 63-74.

Edge, D. (1979). Quantitative measures of communication in science: a critical review. *History of Science,* 17, 102–134.

Frederiksen, L. (2004). Disciplinary determinants of bibliometric impact in Danish industrial research: Collaboration and visibility. *Scientometrics*, 61(2), 253-270.

Kim, M. (2001). A bibliometric analysis of physics publications in Korea, 1994–1998. *Scientometrics*, 50(3), 503-521.

____ . (2005). Korean science and international collaboration, 1995–2000. *Scientometrics*, 63, 321–339.

Price, D. (1963). *Little Science, Big Science*. New York: Columbia University press.

Stokes, T.D., & Hartley, J.A. (1989). Coauthorship, social structure, and influence with specialties, Social Studies of Science, 19, 101–125.

Wagner, C., Brahmakulam, I., Jackson, B., Wong, A., & Yoda, T. (2001). *Science and Technology Collaboration: Building Capacity in Developing Countries*. Callibornia: RAND.

World Bank (2011). *Middle East and North Africa*. Retrieved April 6, 2012, from http://web.worldbank.org/WBSITE/EXTERNAL/COUNTRIES/MENAEXT/0,,menuPK:247619~pagePK:146748~piPK:146812~theSitePK:256299,00.html

KEYWORD-QUERY EXPANSION USING CITATION CLUSTERS FOR PAPER INFORMATION RETRIEVAL

Kiyohiro Yamaguchi[1], Junichiro Mori[2], and Yuya Kajikawa[3]

[1] *kiyoya@gmail.com*
The University of Tokyo, Innovation Policy Research Centre, Yayoi 2-11-16, Bunkyo-ku, Tokyo (Japan)

[2] *jmori@platinum.u-tokyo.ac.jp*
The University of Tokyo, Presidential Endowed Chair for "Platinum Society", Hongo 7-3-1, Bunkyo-ku, Tokyo (Japan)

[3] *kajikawa@mot.titech.ac.jp*
Tokyo Institute University, Graduate School of Innovation Management, Shibaura 3-3-6, Minato-ku, Tokyo (Japan)

Introduction

Query expansion is an old but still active research topic to improve the performance of information retrieval systems by automatically modifying the initial user query. There are some sophisticated approaches using a latent variable model that represents a document as mixtures of topics including Latent semantic analysis (LSA) (Deerwester, et al., 1990), probabilistic LSA (pLSA) (Hofmann, 1999), latent dirichlet allocation (LDA) (Blei et al., 2003). However, because pLSA's generative semantics are not well-defined (Blei et al., 2003), one cannot naturally handle an unseen document. And the number of parameters is affected linearly by the number of seen documents. It has been claimed that these approaches are overfitting the model into the training data. In this paper, we propose a query expansion technique for paper retrieval to integrate topic model and citation network analysis. Citation network analysis is expected to cause moderate topic drift to avoid overfitting.

Data and Methods

Data

We conducted a case study with the dataset of semantic web. We collected data from the Science Citation Index and the Social Sciences Citation Index compiled by the Thomson Reuters (formerly Institute for Scientific Information). Because the whole data is too big, we made small test collections to carry out the experiments. The test collection consists of papers that have the terms semantic and/or web in the keywords. It means that all collected papers have a relation to semantic web to a greater of lesser. We used semantic web as the first query in the experiments. A total of 26,633 papers is counted and the number of citations is 92,441. Only papers published before 2004 (16,638 papers) were used as the training data set, and the other papers published after 2005 were used as the test data set.

Methods

We propose a query expansion for paper information retrieval using not only paper keywords but citation networks to expand a user query. We use a latent semantic space with paper keywords. When we used the latent semantic space derived from the all retrieved data, it tends to be weighted with major terms, which hampers topic expansion and makes expanded corpus similar to the original one. Our main strategy is to divided the dataset into citation clusters and weight the topic model comprehending term features of each cluster. By doing so, it is expected that globally minor but locally important topic can be detected.

First, we built a latent semantic space and obtain an approximation of keyword-paper matrix using LDA. We used lemmatized tokens extracted from paper title and abstract as paper keywords. Next, we built a citation network of the papers and detected clusters from the network using modularity maximization (Newman and Girvan, 2004; Newman, 2004). Then, we obtained the coordinate vectors for each clusters (cluster vectors) using coordinate vectors of their member papers. We use a centroid (arithmetic mean) of vectors as a cluster vector here. The obtained cluster vectors are thought to represent the clusters in the latent semantic space. In this way, we are able to refine the retrieval results in the latent semantic space using detected small research areas extracted by clustering. We used the obtained cluster vectors as a new query and do the same, overall paper retrieval again. This operation, it can be regarded as query expansion, will retrieve other papers in the research areas relevant to the user query. We can repeat this operation if the performance of retrieval increases.

In this paper, we compared two approaches of syntheses; *naïve* and *average* syntheses. We call an expansion using naive synthesis as naive expansion, and an expansion using average synthesis as average expansion. In naive synthesis, we calculate every cosine between a paper and the cluster vectors first. And the highest one is hired as a new similarity of the paper. In the case of average synthesis, we weight cosines between the paper and the cluster vectors by the number of papers in the cluster; multiply by a number of papers in the cluster and divide by the total number of papers. Then, we sum up all of weighted cosines and use it as a new similarity of the paper. We compared these two query expansion approaches to integrate topic model and citation network analysis.

In the experiment, we evaluated how the proposed query expansion technique is capable of retrieving papers in emerging areas related to a query. Here, we built a semantic latent space using the past paper data and did information retrieval and query expansion in the space. If an expanded query is more similar to coordinates of future papers, we can define a coordinate of a future paper in the space built using the past paper data, it means that the query is more relevant to emerging areas. And an area relevant to the expanded query is also relevant to the first query, so these emerging areas are relevant research area to the first query. We build a latent semantic space using a train data set in the collection by LSA and LDA, and evaluate the performance how expanded queries are more similar to papers in a test data set than the first not expanded, only keywords based query.

Results and Discussion

Table 1 shows the R-precision changes of retrieval results at each expansion. We cannot show the all results because of space limitations, we pick up threshold of 0.3 and 0.5 for LDA. Expansion 0 means the result of the original query without expansion, and R-precision of 0 expansion, i.e., the original query, is 49.57%. At the first expansion, the R-precision of average approach of LDA increase than that of the original query. It means that an expansion by the proposed method brings better results than without expansion, while LDA with naïve one have the worth result than the original query. And iterations of the expansion cause topic drift having noisy queries and the performance becomes worse. One can obtain higher performances by only a few naive expansions, but it has a tendency to cause a kind of topic drift to non-relevant papers, which was improved by average approach with citation network analysis.

Conclusion

In the paper, we have proposed the query expansion for paper information retrieval using not only paper keywords but citation networks to expand a query, and evaluated the method using the dataset of semantic web. We showed that the expanded queries have higher similarity to some future papers in that knowledge domain than an original query; the proposed method has the ability to retrieve papers in emerging areas related to the user query. By the experiment results, we conclude that our approach to use both of paper keywords and citation networks works well for paper information retrieval. While we used centroids as representatives in this work, some other methods are known and used as a representative of clusters. In the future work, more experiments with such representatives are necessary. We also need to ensure that retrieval results actually converge after the appropriate number of expansions, or need to find some kind of a stop criterion of expansions.

Table 1. R-Precision (%) changes with the number of query expansion iterations. () is threshold for LDA.

Table	1	2	3
Naïve (0.3)	48.04	30.76	30.89
Average (0.3)	52.17	52.15	52.16
Naïve (0.5)	49.01	44.49	31.61
Average (0.5)	51.78	51.51	51.40

Acknowledgments

We thank Nozomi Nori (Yamaguchi) for her help in the research.

References

Blei D.M., Ng A.Y., and Jordan M.I. (2003) Latent dirichlet allocation. *Journal of Machine Learning Research*, 3, 993–1022.

Deerwester S., Dumais S.T., Furnas G.W., Landauer T.K., and Harshman R. (1990) Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, 41, 391-407.

Hofmann T. (1999) Probabilistic latent semantic indexing. Proceedings of the 22nd Annual International SIGIR Conference on Research and Development in Information Retrieval, pp. 50-57.

Newman M.E.J (2004) Fast algorithm for detecting community structure in networks. *Physical Review E*, 69, 066133.

Newman M.E.J. & Girvan M. (2004) Finding and evaluating community structure in networks, *Physical Review E,* 69, 026133.

KNOWLEDGE COMBINATION FORECASTING BETWEEN DIFFERENT TECHNOLOGICAL FIELDS

Hiroko Nakamura[1], Yuya Kajikawa[2], Ichiro Sakata[3] and Shinji Suzuki[4]

[1] *techhn@mail.ecc.u-tokyo.ac.jp*
University of Tokyo, Center for Aviation Innovation Research, Hongo 7-3-1, 113-8656, Tokyo (Japan)

[2] *kajikawa@mot.titech.ac.jp*
Tokyo Institute University, Graduate School of Innovation Management, Shibaura 3-3-6, 805N, Tokyo (Japan)

[3] *isakata@ipr-ctr.t.u-tokyo.ac.jp*
University of Tokyo, Innovation Policy Research Centre, Yayoi 2-11-16, 113-8656, Tokyo (Japan)

[4] *tshinji@mail.ecc.u-tokyo.ac.jp*
University of Tokyo, Dept of Aeronautics and Astronautics, Hongo 7-3-1, 113-8656, Tokyo (Japan)

Introduction

It is said that innovation is recombination of knowledge and that combining own knowledge to that of different industry and different technology fields have possibility to bring new knowledge creation (ex. Dosi, 1982).

In order to lower the cost of collecting breadth technology knowledge and to accelerate knowledge recombination in a complex technology industry, focusing on one side of breadth activities, searching technologies of other industries, this paper proposes knowledge combination model and discuss how to effectively get the technology field overview of industry and how to identify pairs of technology fields that can be combined between the two industries.

Methodologies

Depth and Breadth Knowledge Combination Model

This paper proposes a knowledge combination model between two technological domains, named the D-B-Combination model (Fig. 1) at research and engineering level and considers two industries of complex system such as automobile and aircraft industries. Figure 1 considers the similarity of any pairs of sub-domains of each technological domain; industry A and of industry B, in the horizon. We assume that similarity of sub-domains can be measured from various sides of characteristics of technological knowledge (ex. problems that the technological knowledge aim to solve, process that the technological knowledge takes to solve a problem) and we also assume that the successful

knowledge combination and the type of recombination between different industries depend on the similarity between the fields.

Sharing knowledge between very similar pair of sub-domains such as the pair of DB-D in Fig. 1 is considered to be important to help deeper understanding and improvement of the technological fields. On the contrary, knowledge export/import of unique technologies from one technology domain to another (pair of DB-T) or knowledge sharing between different technology sub-domains (DB-T) with weak similarity is important to broaden research scope and provides chance of new knowledge creation or transfer/ replacement.

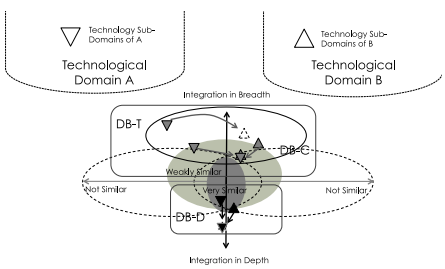


Figure 1. D-B-Combination Model for knowledge integration between two technological domains in depth and breadth

We propose methodologies to measures the similarity of sub-domains of different industries, using patents. And in this paper, as a case study, patents of applicants of major automobile and aircraft makers and suppliers (detail will be at Poster) are analysed. Patent data were retrieved from Thomson Reuters' Thomson Innovation with Derwent World Patents Index (DWPI).

Methodologies

We take three approaches, A1, A2 and A3 and compare the results. Our D-B-Combination model requires data structuring to obtain sub-domains of industries and similarity measurements between sub-domains of two different industries. A1 and A2 take citation analysis approach and A3 takes International Patent Classification (IPC) analysis approach for data structuring. At the citation analysis, the patent and citation data are converted into a non-weighted, non-directed network in which a patent is represented as a node and backward citations to patents as links. The maximum connected component (MC) of the network is extracted and divided into clusters depending on the density of links, using a topological clustering method (Newman, 2004). For similarity measurement, A1 takes cosine similarity measurement approach, A2 takes existing intra-industrial citation tracing approach and A3 takes IPC share similarity comparison approach

Table 1. A1, A2 and A3 result coverage

	A1	A2	A3
Dataset	Automobile: 243,305, Aircraft: 27,989		
Structured Patent Numbers (% over the total)	Automobile MC:60,458 Aircraft MC:8,281 (25.4%)	Combined MC: 69,281 (25.6%)	270,294 (100%)
Identified Cluster	Automobile: 303 Aircraft: 104	420	676
Similarity Analyzed Cluster Numbers	Automobile: 35 Aircraft: 25 (20.6%)	35 (20.5%)	62 (85%)
Highlighted Similarity	48 pairs of clusters	6 clusters	19 sub-classes

Results & Discussion

We discussed, with two senior engineers at Toyota Central R&D Labs, INC. and various level of aeronautic researchers at Japan Aerospace Exploration Agency (JAXA), whether highlighted areas in A1, A2 and A3 identified possible knowledge combination pairs and

provided useful information that support practitioners to create new knowledge.

Table 1 compares A1, A2 and A3 results. For approaches of structuring patents into technology sub-domains, the citation analysis approach limited the structure to the MC so that the coverage of data were much inferior to IPC approach. If comprehensive overview of technological sub-domains in other industries are needed, the IPC approach can satisfy the needs as a list of the technological knowledge better and moreover the overview of the IPC approach can be obtained easily with Microsoft Excel or similar daily software. On the other hand, the citation analysis approach can provide information of technological trend, core knowledge of each field and overview of technological systems.

For measuring similarities between fields (Fig. 2), compared to other two approaches, A1 approach had more possibility to highlight DB-C combinations. It highlighted similarity either in functions (ex. control system) or properties (ex. heat resistance). While it seemed difficult to combine knowledge of same functions but different philosophy, bringing knowledge with different functions but same property or same type of problems together can have possibility of success in knowledge recombination in breadth. Automobile engineers found highlighting such similarity of problems is very useful for them to transfer their technology to other industries and vice versa.

On the other hand, A2 approach highlighted past and emerging knowledge transfer from an industry to another (DB-T combination) such as Avionics, Composite Materials and Fuel Cell and knowledge exchanges in very similar processes (DB-D combination) such as Assembly. The citation analysis enables to structure technology sub-domains as associated parts of systems so that it can support engineers to obtain breadth knowledge related to the technology sub-domains being transferred from or to. It can accelerate broadening scope of projects and adoption of technology.

About A3 approach for measuring similarities, Automobile engineers commented that, even though this approach was simple, some of highlighted areas were interesting to look at.

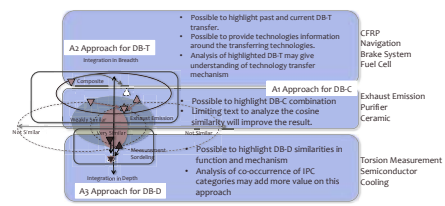


Figure 2. Three methodologies on the D-B-Combination Model

Summary

While it is still needed to improve methodologies, the D-B-Combination model with integration of A1, A2 and A3 methodologies can become an effective innovation designing methodology that allows engineers and product managers to find useful technological knowledge from different industries and explore chances of technological breakthrough in depth or breadth.

References

Dosi G. (1982) Technological Paradigms and Technological Trajectories – a Suggested Interpretation of the Determinants and Directions of Technical Change. *Research Policy*, 11, 147–162.

Newman M.E.J (2004) Fast algorithm for detecting community structure in networks. Physical Review, 69, 066133.

LANGUAGE PREFERENCE IN SOCIOLOGICAL RESEARCH PUBLISHED BY VARIOUS EUROPEAN NATIONALITIES

Moshe Yitzhaki, Yifat Keidar and Nava Rotchild

Dept. of Information Studies; Bar-Ilan University; Ramat-Gan, Israel 52900

Introduction

Recent studies have indicated that scholars of different nationalities are often biased towards their own mother-tongue when using and citing sources (Wood, 1967; Hutchins, 1971; Holmstrom, 1973; Finison-Whittmore, 1975; Chan, 1976; Morgan, 1977; Michel, 1982; Ralph, 1982; Large, 1983; Thorp, 1989; Regaunt, 1994) . However, little research has addressed language preference among non-English-speaking scholars, as compared to their English-speaking colleagues, using additional measures besides the mere percentage of languages cited.

Purpose of the study

The objective of the present study was to assess language self-citation (or language preference) in the field of sociology, gathering empirical data to determine its extent among various nationalities.

Methodology

Forty regular original research articles were drawn from the 1990 and 2010 volumes of 10 sociology journals published in the US, UK, Germany, France, Italy and Spain. Samples included only articles written originally in the language of the journal's publication country and by authors affiliated with institutions there. The references appended to the selected 780 articles were sorted by language, yielding a total of about 25,000 references.

MEASURES USED

Language analysis of cited references in journals of diverse languages reveals the relative use of each language by various nationalities, thus eliciting the 'language self-citation' rate of each group. Clearly, a high rate of self-citation indicates low use of **foreign** language research literature.

Besides the raw figures of 'language self-citation', two more refined measures were employed: the 'relative own-language preference'(ROLP) indicator and the 'Odds Ratio' (Yitzhaki, 1998; Egghe, Rousseau & Yitzhaki, 1999; Bookstein & Yitzhaki, 1999).

Table 1. Frequency Distribution (in %) of languages of cited references in ten sociology journals in 1990

	Percentage of cited references in						
Count. and lang. of citing journals	Eng.	Ger.	French	Span.	Ital.	Other	Total
US	**98.1**	0.8	0.3	0.1	0	0.7	100%
UK	**89.4**	6.6	0.5	0.2	0.1	3.2	100%
Germ.	38.0	**59.7**	2.2	0.2	0	0	100%
France	38.7	1.3	**57.0**	0.6	0.7	1.7	100%
Spain	40.5	0.9	10.3	**43.2**	5.1	0	100%
Italy	30.3	3.0	23.0	4.4	**39.1**	0.2	100%

Findings and discussion

Tables 1 and 2 present the frequency distribution in % of **cited** languages in articles published in the ten journals, grouped by language.

Table 2. Frequency Distribution (in %) of languages of cited references in same ten sociology journals in 2010

	Percentage of cited references in						
Count. and lang. of citing journals	Eng.	Ger.	French	Span.	Ital.	Other	Total
US	**98.6**	0.5	0.5	0.1	0.1	0.3	100%
UK	**95.0**	1.9	0.5	0	0	2.7	100%
Germany	49.5	**50.2**	0.4	0	0	0	100%
France	47.3	0.1	**48.4**	1.6	2.7	0	100%
Spain	47.4	0.7	2.1	**46.2**	3.6	0	100%
Italy	29.4	1.2	12.6	0.4	**56.5**	0	100%

The tables reveal strong 'own-language preference' (OLP) in both periods, varying from one country to another: highest (over 95%) among American and British sociologists, lower among Spanish and Italians ones and lowest among German and French. Besides their own language, German, French and Spanish sociologists heavily cited sources in English, with an increase (10%) in 2010 at the expense of their own language. Spanish sociologists in 1990 still cited sources in French (10%) and Italian (5%) but less in 2010. The Italians had 23% French references in 1990, decreasing to 12.6% in 2010 in favor of their own language. To sum up, among all four non-English nationalities the bulk of cited sources was in English, increasing in 2010, strengthening its position as the 'Lingua Franca' of the new century.

Table 3 estimates the research literature existing worldwide in sociology in both years. To compensate for the English bias in the WOS database, figures of **non-**English items were doubled. Table 4 displays the calculated ROLP indicators and Odds Ratios for each language, assuming that the higher the ratio above 1, the higher the degree of 'self-language preference'.

Table 3. Estimated distribution of languages of sociology reseearch literature (in %) *

Language	1990	2010
English	75.7	74.6
French	11.6	8.3
German	6.6	6.1
Spanish	0.3	0.6
Italian	0.4	0.2
Others	5.4	10.2
Total	**100%**	**100%**

Source: Web of Science.

* Assuming that only **50%** of the **non-**English items are included in this database.

Table 4. Measures of ROLP and 'Odds Ratio' in 1990 and 2010

Country of citing journals	ROLP		Odds Ratio	
	1990	2010	1990	2010
US	1.3	1.3	16.6	24.0
UK	1.2	1.3	2.7	6.5
Germany	9.0	8.2	21.0	15.5
France	4.9	5.8	10.1	10.4
Spain	144.0	77.0	252.8	142.3
Italy	97.7	282.5	159.9	648.1
US	1.3	1.3	16.6	24.0

Both measures indicate that in both periods Spanish and Italian sociologists were much more biased towards their mother-tongue than others. Regarding other nationalities, the two measures differ: according to the ROLP indicators, German authors displayed a higher bias than the French ones, although improving slightly in 2010, while the US and UK scholars had the lowest OLP.

According to the Odds Ratios, however, American and German scholars displayed a higher bias in both periods, compared to British and French ones. The second period shows a slight drop in bias among Germans and rise among the US and UK ones.

Possible Explanations

- Greater availability and accessibility of certain publications than others is probably one reason for OLP, thus enhancing use and citation of own-language material which is usually easier to obtain.
- Most academic libraries in the US and UK do not carry the full array of French and German scholarly journals and monographs, and the same is true for French and German academic institutions regarding English.
- Presumably, leading 'foreign' journals and important monographs are carried by academic libraries in these countries, and most others may be obtained via inter-library loan. Thus, in view of the abovementioned figures, a language barrier or preference may exist alongside the 'immediate availability' factor. Many scholars prefer own-language publications, avoiding the effort of obtaining foreign language material from distant libraries.
- Some of the topics discussed in sociology journals deal with local issues, for which most sources are written in the local language.

Conclusions

1. To determine the extent of the 'relative own-language preference' among scholars in certain fields, one may relate the 'language self-citation' rate of a each language-group, to some estimate of this language share in the field research literature.
2. In sociology both of the measures used revealed a similar state for Spanish and Italian sociologists, a differing one for other nationalities and the lowest degree of mother-tongue bias among the British scholars.

References

Bookstein, A and Yitzhaki, M. (1999). Own-language preference: a new measure of 'relative language self-citation'. *Scientometrics,* 46, 337-348.

Chan, G.K.L. (1976). The foreign language barrier in science and technology. *International Library Review* 8, 317-325.

Egghe, L, Rousseau, Yitzhaki, M. (1999). The 'own-language preference': measures of relative language self-citation". *Scientometrics* 45, 217-232.

Finison, L.J. and Whittemore, C.L. (1975). Linguistic isolation of American social psychology. *American Psychologist* 30, 513-516.

Holmstrom, J.E. (1973). The foreign language barrier. In J. Sherrod, & A. Hodina (Eds.). *Reader in Science Information*. Washington, D.C.: Microcard, 94-103.

Hutchins, W.J. a.o. (1971). *The Language Barrier*. University of Sheffield.

Large, J.A. (1983). The Foreign-Language Barrier; Problems in Scientific Communication. London: Andre Deutsch.

Michel, J. (1982). Linguistic and political barriers in the international transfer of information in science and technology. *Journal of Information Science* 5, 131-135.

Morgan, B.A. (1977). National bias in reference citation in English language agricultural engineering journals. *IAALD Quarterly Bulletin,* 22(3-4), 60-64.

Ralph, A. (1982). Language and information retrieval in the social sciences. *Aslib Proceedings* 34, 394-405.

Regaunt, S. (1994). English as lingua franca in geological scientific

research; a bibliometric study. *Scientometrics* 29, 335-351.

Thorp, R.G. a.o. (1989). The foreign language barrier: a study among pharmaceutical research workers. *Journal of Information Science.* 14, 17-24.

Wood, D.N. (1967). The foreign-language problem facing scientists and technologists in the UK-report of a recent survey. *Journal of Documentation*. 23, 117-130.

Yitzhaki, M. (1988). The language barrier in the humanities: the case of biblical studies. *INFORMETRICS 87/88.* Amsterdam: Elsevier, 301-314.

Yitzhaki, M. (1998). The 'language preference' in sociology: Measures of 'language self-citation', 'relative own-language preference indicator' and 'mutual use of languages'. *Scientometrics,* 41, 243-254.

LEADERS AND PARTNERS IN INTERNATIONAL COLLABORATION AND THEIR INFLUENCE ON RESEARCH IMPACT

Borja González-Albo, Javier Aparicio, Luz Moreno, María Bordons

borja.gonzalezalbo@cchs.csic.es
IEDCYT, CCHS, Spanish National Research Council (CSIC), Albasanz 26-28, Madrid (Spain)

Introduction

A positive influence of international collaboration on the impact of research has been extensively described (see for example, Glänzel 2001). Different underlying reasons may interact. On the one hand, the higher number of authors and institutions described for this type of collaboration can be a key issue, since it may include higher variety of points of view which may lead to higher creativity and innovation (Reagans and Zuckerman 2001). On the other hand, teams can benefit especially from the knowledge, infrastructures and prestige of scientifically advanced partners.

Focusing on Spain, a higher impact for internationally co-authored articles as compared to the total output of the country has been reported in the literature (Bordons, Aparicio and Costas, 2013). The interest of considering additional factors such as the type of collaborating country (Gazni et al. 2012) and which partner takes the lead in the research (Van Leeuwen and Tijssen, 2007) has also been suggested, but as far as we know it has not been globally addressed before.

Objectives

This paper analyses the internationally co-authored papers of Spain to explore to what extent the type of partner (high or low R&D intensive country) and its role in the collaboration (leadership as measured by first authorship) may influence the impact of research.

The following questions are addressed: a) In which proportion does Spain collaborate with high and low R&D intensive countries? Are there any differences by fields? b) Is there any relationship between type of partner and impact of research? Specifically, are papers co-authored with high R&D intensive countries published in more prestigious journals? Do they receive more citations? c) Are papers leaded by high R&D intensive countries more heavily rewarded through citations?

These questions are studied in the total production of Spain during a two year period and differences by fields are explored.

Methods

Scientific publications of Spain in 2008-2009 as covered by the Web of Science database are analyzed. Only articles, proceedings papers and reviews are considered ("articles").

The study focuses on bilateral collaboration. Only publications with 2 addresses, one from Spain and the other from another country are taken into account, to avoid potential confusion derived from papers with multi-lateral collaboration.

Countries are classified in three classes according to their level of commitment with research and development activities as measured by their gross domestic expenditures in R&D as a percentage of their gross domestic product (%GERD/GDP) (source: World Bank 2012). Spain devoted 1.35% of its GDP to R&D in year 2008. Countries are classified as high RD (%GERD/GDP higher than 1.55); low RD (%GERD/GDP below 1.15) and similar to Spain (1.15>%GERD/GDP>1.55).

The impact of research is measured through: a) journal normalized position (JNP), which is a measure of the prestige of journals, since it takes into account the position of a journal in its field considering all journals in descending order of impact factor (Journal Citation Reports); b) citations per paper, which consider citations received by publications within a 3-year citation window. To make inter-field comparisons possible the citation counts of papers are normalized with respect to the average citation rate of Spain in the discipline of the corresponding journal and in the same period of time (RCR).

"First authorship" is considered as an indication of scientific leadership (Van Leeuwen and Tijssen, 2007).

Publications are assigned to disciplines according to the Web of Science classification of journals into subfields. Disciplines are grouped into ten broad areas.

Results

The scientific output of Spain in WoS during 2008-2009 accounts for 84,366 articles; and 41% of them includes at least one foreign address. Bilateral collaboration is present in 9,800 articles (28.3 % of internationally co-authored articles). Collaboration with high RD countries (60%) predominates over that with low RD countries (25%), although small differences by broad areas are observed. Spain is in the first address in 56% of the publications, but this percentage ranges from 53% in Clinical Medicine to 60% in Agriculture.

Papers co-authored with high RD countries obtain higher impact (JNP and relative citation rate) than those co-authored with low RD countries for the total country and in all areas -although in Mathematics the differences are limited to JNP- (Figure 1). Moreover, among the most cited papers (21% of the papers which obtained a RCR equal or higher than 1.5), high RD partners are over-represented. This holds for the total production and also for each of the broad areas except for Mathematics.

Do Spanish articles obtain higher impact when a high RD country leads the research (first-authorship)? This is confirmed only in Physics, while in the rest of the areas no significant differences in impact (JNP or relative citation rate) are observed according to the position of the Spanish centre in the address.

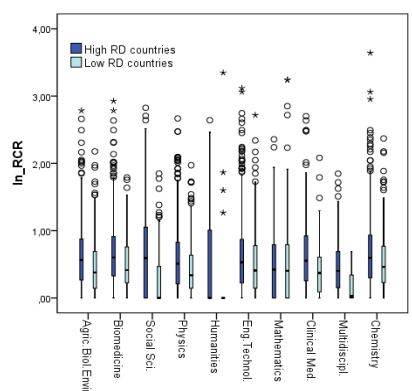


Figure 1. RCR by type of partner and broad area

Focusing on Spanish articles internationally co-authored with low RD countries, a higher impact is observed for articles with a Spanish first address

in Agriculture, Biomedicine and Chemistry (higher JNP and RCR); in Engineering (higher JNP) and in Multidisciplinary (higher RCR).

Conclusions

The fact that Spain collaborates more often with high than with low RD countries is not surprising, since scientifically advanced countries are more attractive partners (Gazni, Sugimoto and Didegah, 2012) and this type of collaboration is supposed to yield higher benefits for research teams. Our results confirm this assumption, since Spanish papers co-authored with high RD countries are published in more prestigious journals, receive more citations and are more likely to be amongst the most cited papers than those co-authored with low RD countries. Similar results were described for a particular discipline in Spain (Bordons, Aparicio and Costas, 2013) and for different areas in Latin-American countries (De Filippo, Aparicio and Gómez, 2009).

However, an interesting result is that when collaborating with a high RD country, Spain may lead the research as often –or in some areas more often– than the other country and no differences in impact according to the position of partners in the address field are observed, that is, foreign leadership does not originate higher impact. On the contrary, in the set of papers co-authored with low RD countries, the research leaded by Spain shows a higher impact in four areas. A possible explanation is that the collaboration with high RD countries is more symmetric (Kim, 2006) while a more asymmetric nature may characterise the one conducted with low RD countries, in which Spain might participate as a "strong" partner. Finally, the special behaviour of Mathematics is discussed.

Acknowledgments

This research has been supported by the Spanish National Plan R&D (Research project CSO2008-06310).

References

Bordons, M.; Aparicio, J.; Costas, R. (2013). Heterogeneity of collaboration and its relationship with research impact in a biomedical field. *Scientometrics*. DOI 10.1007/s11192-012-0890-7.

De Filippo, D.; Aparicio, J.; Gómez, I. (2009). Measuring the benefits of International collaboration. A case study of the relationship between Latin-American and European countries. B.Larsen and J.Leta, Eds. *Proceedings of the 12th International Conference of the ISSI*. Brasil: BIREMEPPAHO/WHO and Federal University of Rio de Janeiro, 2009. Pp. 920-921.

Gazni, A.; Sugimoto, C.R.; Didegah, F. (2012). Mapping world scientific collaboration: authors, institutions and countries. *Journal of the American Society for Information Science and Technolog,y* 63(2): 323-335.

Glänzel, W. (2001). National characteristics in international scientific co-authorship relations. *Scientometrics*, 51 (1), 69–115.

Gorraiz, J.; Reimann, R.; Gumpenberger, C. (2012). The importance of bilateral and multilateral differentiation in the assessment of international collaboration -a case study for Austria and six countries. *Scientometrics,* 91(2):417-433.

Reagans, R.; Zuckerman, E.W. (2001). Diversity and productivity: the social capital of corporate R&D teams. *Organization Science,* 12(4): 502-517.

Van Leeuwen, T.N.; Tijssen, R. (2007). Strength and weakness of national science systems. A bibliometric analysis through cooperation patterns. In: Torres-Salinas, D. and Moed, H.F. Ed. *Proceedings of the 11th International Conference of the ISSI.* Madrid: CINDOC-CSIC. Pp. 469-479.

MEASURING INTERDISCIPLINARITY OF RESEARCH GRANT APPLICATIONS. AN INDICATOR DEVELOPED TO MODEL THIS SELECTION CRITERION IN THE ERC'S PEER-REVIEW PROCESS

Ivana Roche[1], Dominique Besagni[1], Claire François[1], Marianne Hörlesberger[2], Edgar Schiebel[2], Dirk Holste[2]

[1] *ivana.roche@inist.fr; dominique.besagni@inist.fr; claire.francois@inist.fr*
CNRS, Institut de l'Information Scientifique et Technique, UPS 76, 2 allée du Parc de Brabois, Vandœuvre-lès-Nancy, F-54519 Cedex (France)

[2] *marianne.hoerlesberger@ait.ac.at; edgar.schiebel@ait.ac.at; dirk.holste@ait.ac.at*
AIT Austrian Institute of Technology GmbH, Donau-City-Strasse 1, 1220 Vienna (Austria)

Introduction

Scientific experts having to evaluate the interest of research projects submitted for grant to the European Research Council (ERC) must lean on the four "frontier research" criteria among which: *"... it pursues questions irrespective of established disciplinary boundaries, involves multi-, inter- or trans-disciplinary research that brings together researchers from different disciplinary backgrounds, with different theoretical and conceptual approaches, techniques, methodologies and instrumentation, perhaps even different goals and motivations*", according to the EC's High Level Expert Group report (2005) assertions and that we defined in our DBF project (Hörlesberger *et al.*, 2013) as interdisciplinarity.

Background

Interdisciplinarity is used as a proxy to infer self-consistently the presence and proportions of characteristic terminology associated with individual ERC panels, thereby revealing the intra or inter-panel character of a proposal. It was built upon a previously successfully tested approach that the frequency of occurrence and distribution of research field specific keywords can classify and characterize research fields. While the core of the approach has been retained, the computation has been adopted and fine-tuned to the grant scheme under study. The underlying basic hypothesis is that the larger the proportion of inter-panel keywords, the more interdisciplinary is the proposal. To this end, each keyword is labeled according to its statistical frequency across all panels, filters are applied to distinguish relevant from irrelevant keywords, and the tallying of keywords with their assigned panels is assessed to classify each proposal with respect to its share of inter-panel keywords.

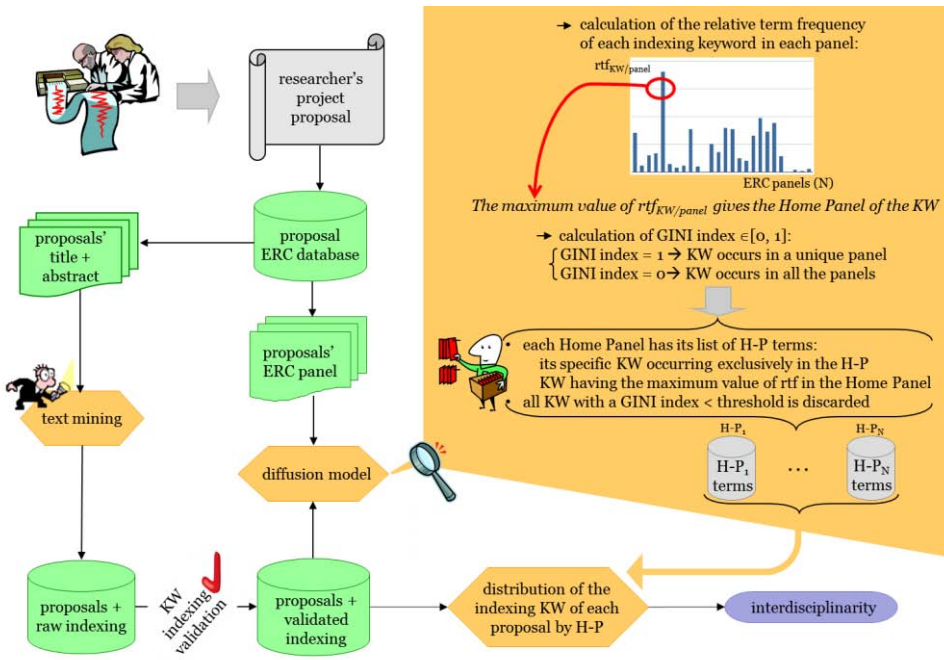


Figure 1. Methodological schema of the evaluation process of the interdisciplinarity of research grant applications

Data and Methodology

Primary data are directly obtained from the documents supplied by the researcher. Each project proposal is indexed by keywords (KW) extracted from its textual information. After validation, we apply the diffusion model approach. First, the so-called Home Panel (H-P) terms are defined. We assumed KW which are specific for a panel occurred with a higher probability in that panel rather than in others. This probability is defined by the frequency of one KW in a panel divided by the number of proposals in this panel, namely the relative term frequency ($rtf_{KW/panel}$). Then, for a KW, we calculate its $rtf_{KW/panel}$ in each panel and the panel with the highest probability is declared to be its H-P. So, we obtain for each H-P the list of its H-P terms. Therefore the complete terminology associated to a panel consists of the union of its H-P term list and the set of terms imported from the other panels and we refer to as "abroad terms".

In order to make the lists of H-P terms more consistent, we use the Gini index, a measure of the inequality of a distribution, varying from *0* to *1*, a value of *0* expressing total equality and a value of *1* maximal inequality. It is commonly used in economics as a measure of inequality of the income or wealth. It is, in this study, employed as a measure of the dispersion of a KW. A Gini index of *1* tells us that the KW is very specifically limited to only one panel. Conversely, a Gini index equal to *0* means a completely uniform distribution and indicates that the term occurs in all the considered panels. We define a cut-off threshold to discard KW which dispersion will be considered excessive.

Finally, the tallying of KW with their newly assigned H-P is assessed to

calculate an index for the share of abroad terms in the list of KW representing the content of each project proposal. The underlying assumption is that the larger the proportion of abroad terms, the more interdisciplinary a project proposal is.

The Case Study

We applied our methodology to a case study coming from project proposals submitted to the ERC 2009 Starting Grant Call.

Table 1. Distribution of successful and non-successful proposals submitted to the ERC 2009 Starting Grant Call

	ERC StG 2009	Dataset (6 panels)
Proposals	2.503	198
Successful	244	41
Non-successful	2.259	157

Among the 19 ERC panels representing Life Sciences (LS) and Mathematics, Physics, Chemistry, Engineering & Earth Sciences (PE), we chose 6 panels with a balance between LS and PE as well as between basic and applied fields.

The analysis shows the results for ERC panel PE1 (Mathematics and Mathematical foundations) which received 39 submissions, 11 of them successful. The value of Interdisciplinarity in this example is the share of abroad terms calculated with a Gini index cut-off threshold of 0.1.

In the DBF project, we introduced a statistical discrete choice model (DCM) to estimate the decision probability for a proposal to be accepted on the basis of measured attributes of "frontier research" (Scherngell *et al.*, 2013).

We studied in a proof-of-concept approach the influence of those attributes on the success probability and conducted an initial analysis of the *ex-post* comparison between the indicator-based scientometric evaluation and the empirical peer-review process. Interdisciplinarity is one of the five indicators we developed in the context of this project and it is an element that proves to influence significantly the selection probability of the projects.

The results obtained with Interdisciplinarity are encouraging and we are experimenting with it on a different dataset from e-Corda.

We also have to introduce another dimension to that indicator: the diversity of the sources (H-P) of abroad terms, on top of just counting them.

Acknowledgments

Work accomplished in the context of the DBF (Development and Verification of a Bibliometric Model for the Identification of Frontier Research) project within the European Research Council's CSA of the EU's 7th Framework Programme.

References

Hörlesberger M., Roche I., Besagni D., Scherngell T., Francois C., Cuxac P., Schiebel E., Zitt M., and Holste D. (2013). A concept for inferring "frontier research" in grant proposals, Scientometrics (to appear).

Scherngell T., Roche I., Hörlesberger M., Besagni D., Züger M.-E., and Holste D. (2013). Initial comparative analysis of model and peer-review process for ERC starting grant proposals, Research Evaluation (submitted).

MEASURING THE QUALITY OF ACADEMIC MENTORING

Jongwook Lee

jl12b@my.fsu.edu
Florida State University, College of Communication and Information, Tallahassee, FL 32306 (United States)

Introduction

A mentor is a critical factor for the career and psychosocial development of a mentee (Kram, 1983). In academic context, faculty members may be mentors to their students. Sugimoto (2012a) confirmed that faculty advisors (i.e., dissertation advisors) can be mentors to their doctoral students in library and information science (LIS). She also applied Kram's mentoring framework to LIS doctoral education and identified the importance of the relationships between advisors and their doctoral students for successful doctoral studies (Sugimoto, 2012b). Despite the significance of academic mentoring in doctoral education, there has been little research on measuring the quality and impact of academic mentoring. Several attempts have been made to measure the academic mentoring of faculty quantitatively based on the number of doctoral mentorship pairings (Marchionini, Solomon, Davis, & Russell, 2006; Sugimoto, 2006). This study aims to propose a method of measuring the quality of academic mentoring.

Background

Marchionini, Solomon, Davis, and Russell (2006) describe academic mentoring as an activity that covers the research, teaching, and service activities of faculty. Their study operationalizes mentoring as "service as doctoral advisor and service on doctoral dissertation committees" (p. 481). They present the MPACT indicators for mentors based on this operational definition. Sugimoto (2012a) supports this study by examining the features of academic mentoring in LIS. She surveyed 354 tenured professors in LIS programs that grant a doctoral degree and 294 assistant professors from the ALA-accredited LIS programs. Of the respondents to the questionnaires, 33 advisors and 23 advisees were interviewed. The research found that advisors could be regarded as mentors. Furthermore, Sugimoto et al. (2008) found a statistically insignificant correlation between the MPACT values and faculty citation counts and claimed that the MPACT values (i.e., sum of the number of doctoral student advising as chairs and as committee members) can show a different aspect of faculty scholarship. The limitation of MPACT score is, however, that it shows only the quantitative aspect of mentoring, rather than the quality of mentoring.

Some research has attempted to examine the effect of mentoring in graduate education on the research productivity of mentees. Paglis, Green, and Bauer (2006) measured research productivity of doctoral students for five and half years as they began their doctoral studies. The findings show that research collaboration with advisers influences

the research productivity of advisees positively, and "psychological mentoring" improved the research self-efficacy of advisees. That is, the positive relationship between advisees' research productivity and advisors' mentoring was observable. Tenenbaum, Crosby, and Gliner (2001) studied the relationship of mentoring, student satisfaction, and scholarly productivity in graduate school. This study demonstrates that "instrumental mentoring" can improve advisees' research productivity, and "psychosocial help" can increase the satisfaction of advisees with their advisors. In a more recent study, Muschallik and Pull (2012) showed that effective mentoring facilitates the transfer of knowledge and skills from mentors and mentees, which will result in increasing the research productivity of mentees.

Method

Given that the research productivity of doctoral students is influenced by the academic mentoring provided by faculty advisors, this study used advisees' research productivity (i.e., publication and citation count) for assessing mentoring quality of advisors.

Population

The population of this pilot study is twelve tenure-track LIS faculty members (i.e., full and associate) who are employed by the Florida State University as of January 2013. Assistant professors are not included in this population because they have not had many opportunities to advise students. Two full faculty members whose backgrounds are not LIS were also excluded because the MPACT database used for collecting doctoral mentorship information does not cover all disciplines completely while it has the complete information of LIS.

Data Collection

The faculty list is collected from the LIS website. Advisee lists of faculty members are gathered from the MPACT site (http://www.ibiblio.org/mpact/). Once the lists of advisees are made, publication and citation data of each advisee were manually collected from Web of Science by searching with their names. The author initially searched SSCI database and collected bibliographic and affiliation information of advisees. After that, the author used "Author Search" function in order to expand the scope of database to SCI and A&HCI using affiliation information of advisees.

Table 1. Faculty rankings based on the quantity and quality of academic mentoring

Faculty	#1	#2	#3	R1	R2	R3
A1	29	96	354	1	2	2
A2	25	80	344	2	3	3
A3	15	304	1995	3	1	1
A4	10	17	114	4	5	4
A5	6	5	19	5	7	6
A6	6	6	11	5	6	7
A7	5	19	45	7	4	5
A8	4	3	1	8	8	8
A9	3	0	0	9	10	10
A10	3	1	0	9	9	9
A11	2	0	0	11	11	11
A12	0	0	0	12	12	12

#1: Number of doctoral dissertation advising as chair and committee (MPACT values)
#2: Publication count of mentees
#3: Citation count of mentees
R1: Ranking by #1
R2: Ranking by #2
R3: Ranking by #3

Findings

Faculty members are ranked by three criteria: (a) number of advisees; (b) publication count of advisees; (c) citation count of advisees (see Table 1). The ranking changes were found among three measures. For example, the ranking of A3 is improved from 3 to 1 as the publication and citation counts of

advisees were considered. In addition, A7 is ranked 7th, 4th and 5th by three criteria respectively.

Limitations and Future Work

This pilot study implies the need for assessing both the quantity and quality of academic mentoring. However, this study has limitations in several aspects. First, the size of population is too small to discover meaningful patterns that can be generalized. To deal with this limitation, future research will include faculty members from other LIS schools in the United States. In addition, although some studies demonstrated the importance of using multiple databases in evaluative bibliometric studies (Meho & Yang, 2007), this study is limited to the Web of Science in collecting publication and citation data. The future study will attempt to analyse the data gathered from Web of Science and Scope. Finally, other variables such as research areas or prior research skills of advisees may influence their research productivity. To strengthen these limitations, research interests of advisors and advisees as well as the characteristics of advisees need to be identified.

Acknowledgments

I would like to thank to my mentors, Dr. Kathy Burnett and Dr. Kiduk Yang, for giving me valuable advice.

References

Kram, K. E. (1983). Phases of the mentor relationship. *Academy of Management Journal, 26*(4), 608-625.

Marchionini, G., Solomon, P., Davis, C., & Russell, T. (2006). Information and library science MPACT: A preliminary analysis. *Library & Information Science Research, 28*(4), 480-500. doi: 10.1016/j.lisr.2006.04.001

Meho, L. I., & Yang, K. (2007). Impact of data sources on citation counts and rankings of LIS faculty: Web of Science versus Scopus and Google Scholar. *Journal of the American Society for Information Science and Technology, 58*(13), 2105-2125.

Paglis, L. L., Green, S. G., & Bauer, T. N. (2006). Does adviser mentoring add value? A longitudinal study of mentoring and doctoral student outcomes. *Research in Higher Education, 47*(4), 451-476.

Sugimoto, C. R., Russell, T. G., Meho, L. I., & Marchionini, G. (2008). MPACT and citation impact: Two sides of the same scholarly coin? *Library & Information Science Research, 30*(4), 273-281. doi: 10.1016/j.lisr.2008.04.005

Sugimoto, C. R. (2012a). Are you my mentor? Identifying mentors and their roles in LIS doctoral education. *Journal of Education for Library and Information Science, 53*(1), 2-19.

Sugimoto, C. R. (2012b). Initiation, cultivation, separation and redefinition: Application of Kram's mentoring framework to doctoral education in information and library science. *Journal of Education for Library and Information Science, 53*(2), 98-114.

Tenenbaum, H. R., Crosby, F. J., & Gliner, M. D. (2001). Mentoring relationships in graduate school. *Journal of Vocational Behavior, 59*, 326-341.

A MODEL BASED ON BIBLIOMETRIC INDICATORS: THE PREDICTIVE POWER

Elizabeth S. Vieira[1], José A. S. Cabral[2], José A.N.F. Gomes[1]

[1] *jfgomes@fc.up.pt*
REQUIMTE/Departamento de Química e Bioquímica, Faculdade de Ciências, Universidade do Porto, Rua do Campo Alegre, 687, 4169-007 Porto, Portugal

[2] *jacabral@fe.up.pt*
INESC-TEC, Faculdade de Engenharia, Universidade do Porto, Rua Dr. Roberto Frias, s/n, 4200-465 Porto, Portugal

Introduction

Bibliometric indicators have been widely used for assessing the scientific performance of a given research body. The design of indicators has attracted a lot of attention in the last few years as national authorities, funding bodies or institutional leaders show a growing interest in indicators that can rate the performance of their institutions. Nowadays, several countries use a combination of peer review and bibliometric indicators to assess the research performance of higher education institutions and to allocate funding. Peer review is still the gold standard of research evaluation but the pressure for more frequent and extensive assessment of the performance of researchers, research groups and research institutions makes bibliometry attractive. It is therefore very important to benchmark bibliometric indicators against traditional peer assessment in real situations. Some studies have been carried out in recent years with the main goal of finding a correlation between the two methodologies. These studies considered the judgments of peer-review at several levels: 1) national level (Franceschet and Costantini 2011); 2) research programs, research groups, departments (Aksnes and Taxt 2004; Van Raan 2006) and 3) at individual level (Bornmann and Daniel 2006, 2007; Bornmann, Wallon and Ledin 2008). At the individual level the application of bibliometric indicators is more complex and more care is needed in the data treatment. However, the application of such indicators is feasible as a few studies made in the last few years (Bornmann et al. 2006, 2007; Bornmann et al. 2008) suggest.

Here we consider the peer decisions in a collection of academic job openings in Portuguese universities and look for ways to design a model based on bibliometric indicators that may follow these results. Considering the high weight that the scientific performance has in these selection processes, we looked for a set of indicators that we suggest may be thought to be implicit in the judgment by the peers. We are not suggesting that the ranking of candidates in this type of job openings can be made without recourse to peers. However, we consider that if a definition of a model based on bibliometric indicators is feasible, this complementary instrument can help peers in the evaluation of the applicants and the formulation of the final decisions. The predictive power of the model was explored aiming at finding how far the model can predict

the peer decisions on the academic openings.

Methodology

The data set consists of 27 contests with a total of 171 researchers. The number of applicants varies among contests (2-11). A set of 12 indicators were considered based on the list of the candidates' publications indexed in the ISI Web of Science. The indicators used were:

NDF – Number of documents. Each document was divided by the number of authors, (1/N, N being the total number of authors in each document).

NIR – The normalized impact indicator for researchers. This indicator gives the mean number of normalized citations per document taking into account the different culture of citation of each subject category.

h_{nf} – This indicator is calculated in a way similar to the *h* index, but considers the different cultures of citation of each subject category and the number of authors per publication.

HCD – Percentage of highly cited documents. This indicator considers the percentage of documents of a given researcher that are in the Top 10% most cited Portuguese documents.

DIC – Percentage of documents with collaboration. This indicator refers to documents with at least one international collaboration.

CD - Percentage of citing documents. This indicator does not consider those citing documents that belong to the researcher.

DNC – Percentage of documents not cited.

SNIP – Source normalized impact indicator (Moed 2010).

SJR – SCImago journal rank (Gonzalez-Pereira, Guerrero-Bote and Moya-Anegon 2010).

Q1 – High prestige journals (SCImago Institutions Rankings).

NI – Normalized impact (SCImago Institutions Rankings).

NAm – Number of normalized authors per document.

The rank ordered logistic regression (ROLR) was applied to our data in order to define the model. After defining the model, the predictive power was studied. Initially this was done determining the number of times the candidate placed in the first position by peers was that with the highest probability of being chosen first. Then we have studied the model forecasts and the peer selection using all pairs of applicants.

Results

The application of the ROLR lead to the following model:

$$P_i = \frac{e^{0.40h_{nf_i}+0.032HCD_i}}{\sum_j e^{0.40h_{nf_i}+0.032HCD_i}}$$

where P_i is the probability of the applicant *i* to be selected in first place by peers.

Table 1. First place, correctly, predicted by the model.

	Model	Random estimate
Correct prediction of the first place	52%	23%

Once we have the model defined, we determined the percentage of cases where the predictions given by the model are coincident with the peer judgments. For the set of openings considered, it was determined the number of times the candidate placed in the first position by peers was that with the highest probability of being chosen first. The result is compared with a

situation without information about the applicants. The results are presented in Table 1.

The model defined allows us to go further in the assessment of the predictive power. It is possible to evaluate the predictions given by the model using pairs of applicants. In the data set considered, 426 pairs are available and the model predicts, correctly, 75% of the pairs. The Monte Carlo's simulator was used to obtain the probability distribution function as presented in Figure 1.

The fitting of a normal distribution to the data gives a probability distribution function that is skewed and platykurtic (-0.047 and -0.029 respectively). The results of the fit suggest that the model, on average, predicts correctly 72%±2% of the pairs. The percentage predicted by the model (75%) is in the range [μ+2σ]. The probability of predicting correctly more than 75% of the pairs is just 7.35%.

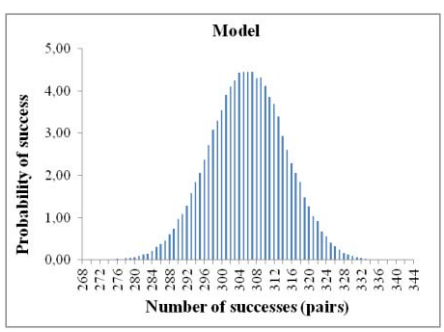


Figure 1. Probability distribution function obtained using Monte Carlo's simulation.

Acknowledgments

Elizabeth Vieira wishes to acknowledge the financial support from FCT, Portugal, through a grant Nº SFRH/BD/75190/2010.

References

Aksnes, D. W. & Taxt, R. E. (2004). Peer reviews and bibliometric indicators: A comparative study at a Norwegian university. *Research Evaluation*, 13(1), 33-41.

Bornmann, L. & Daniel, H. D. (2006). Selecting scientific excellence through committee peer review - a citation analysis of publications previously published to approval or rejection of post-doctoral research fellowship applicants. *Scientometrics*, 68(3), 427-440.

Bornmann, L. & Daniel, H. D. (2007). Convergent validation of peer review decisions using the h index - extent of and reasons for type i and type ii errors. *Journal of Informetrics*, 1(3), 204-213.

Bornmann, L. Wallon, G. & Ledin, A. (2008). Does the committee peer review select the best applicants for funding? An investigation of the selection process for two european molecular biology organization programmes. *Plos One*, 3(10).

Franceschet, M. & Costantini, A. (2011). The first italian research assessment exercise: A bibliometric perspective. *Journal of Informetrics*, 5(2), 275-291.

Gonzalez-Pereira, B. Guerrero-Bote, V. P. & Moya-Anegon, F. (2010). A new approach to the metric of journals' scientific prestige: The sjr indicator. *Journal of Informetrics*, 4(3), 379-391.

Moed, H. F. (2010). Measuring contextual citation impact of scientific journals. *Journal of Informetrics*, 4(3), 265-277.

Van Raan, A. F. J. (2006). Comparison of the hirsch-index with standard bibliometric indicators and with peer judgment for 147 chemistry research groups. *Scientometrics*, 67(3), 491-502.

MONITORING OF INDIAN RESEARCH PAPERS: ON THE BASIS OF MAJOR GLOBAL SECONDARY SERVICES

Divya Srivastava[1], Arvind Singh Kushwah[2], Mona Gupta[3]

drdivya.srivastava@gmail.com[1] , arvindsinghmca@yahoo.com [2], Gmona7@gmail.com [3]
Scientist, Indian Council of Medical Research, Ansari Nagar, New Delhi 110029, India

Background

Science and technology pursuit has been a major planning objective of the country, identified on purpose to initiate, advance and accelerate national development in all sectors of the economy. Consequent upon this policy initiative, India has been able to usher significant growth in its capacity and capability in basic, applied, and developmental research in science and technology. Its S&T infrastructure has grown large, comprising of more than 300 universities, 400 research laboratories, 13 institutes of national importance. It is desirable that India comes out with a program to measure and monitor its performance in S&T on a regular basis. This task inevitably requires building appropriate indicators of S&T performance. Besides, indicators are required for depicting how Indian science is being covered by major secondary services. It is within such a context that the present study has been done. India has the third largest pool of researchers in the world, after the USA and the CIS (Commonwealth of Independent States). However, the quality of research output does not reflect this numerical strength. The literature has pointed that that in the present situation a lot of papers gets missed out and visibility and impact of Indian research efforts gets affected adversely. There are studies depicting that the total Indian share in world scientific output appearing in a scholarly journal covered by the Science Citation Index (SCI) between 1981 and 1995 has declined by 32%, almost double the 17% decline estimated in World Science Report, on the basis of SCI data collected for 1982 and 1993. Indian scientists were responsible for close to half of the third world's publication output in 1973 and India with 7888 papers was the eighth largest publishing nation in the world. Now India's rank has slid to fifteenth in 1998 and 2000. The number of papers published (as seen from SCI) over two decades shows similar trend. The number of papers published in 1980, 1985, 1990, 1995, 2000, 2005 was 14,983, 11,222, 10,103, 11,084, 12,127 and 25,102 respectively. This clearly indicates that number of research papers published by Indian scientists is around 10,000-25,102^{+} during the last two decades. To test this hypothesis we have downloaded the papers having 'India' in its address field and appeared in a journal covered by SCI for the period of 2005-2009. There were total 1,74,403 papers indexed in SCI being published by the over 40 CSIR affiliated labs, 31 ICMR Institutes, Institutes belonging to DST, ICAR and the rest were from the university/ colleges sectors. The percentage contribution was 1.78, 26.42 and 29.25 from Colleges, Universities and the research Institutions.

Most of the institutions publish their papers in non SCI journals, therefore, there is a need that, total Indian research papers should be computed on the basis of papers from major global services along with papers published by Indian authors in foreign journals to generate need based Indian National Science Indicators/ reports for the national, and geographical productivity as well as subject wise productivity depicting the trend of research papers. With this background, the present study would provide a consolidated and comprehensive sound database on amount of work/research done in the field in India and will facilitate a quick access to various Indicators.

Methodology

In the present study, the papers have been extracted from selected global secondary services *eg*, INSPEC, Tropical Diseases Bulletin (Online Service), CABI, AGRIS and Indian Science Abstracts by Indian authors. The data has been analyzed separately for each database for computing trend of papers being covered by that particular secondary service along with the trend of coverage of journals and the pattern of subject area over the years (2005-2009). A study has also been done to compute total publications during the period of the study which are coming out from different cities and states of India.

Observations & Findings

The objectives of the study are threefold: assessing the contribution of different cities and states to mainstream scientific literature in different disciplines of science and technology during 2005-2009; identifying most prolific City, Institutions and their contribution in different broad disciplines of science and technology. These outputs are used to understand growing Indian capacities and potential in different fields of science and technology. In the period 2005 - 2009, the Indian S&T output as reflected in all the databases, is skewed. In the year of 2007 a significant growth has been observed. The Graph 1 shows the percentage of papers produced by Indian scientists in the five years, which has been covered by all the three databases. Despite the limitations in funding for science and technology, the contribution of Indian scientists to the world's scientific output increased during the last five years. India established a large number of institutions in the recent period which is yielding currently large number of research papers.

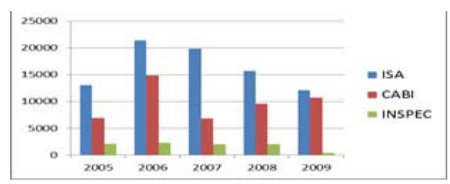
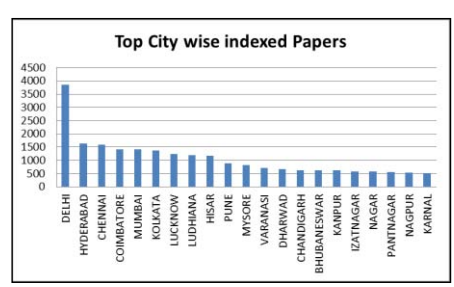


Figure 1. Year wise Total Papers covered by ISA, CABI & INSPEC.

An analysis has been done to compute the productivity of Indian cities and the Indian States also on the basis of papers indexed in the secondary services *viz* ISA, INSPEC and CABI. A total of 666 Indian Cities contributed all these papers. The top most city was Delhi followed by Hyderbad, Chennai, Coimbatore and Mumbai.



The data was analyzed to compute the productivity of Indian States also. All the 30 Indian States produced papers which were covered by the services under study. The top most state with (more than 1000 Papers) maximum indexed paper was Uttar Pradesh followed by Tamil Nadu, Maharashtra, Delhi, Karnataka, Andhra Pradesh, West Bengal, Haryana, Rajasthan, Punjab, Madhya Pradesh, Gujarat and Uttarakhand.

Coverage of Science and Technology Journals by Secondary Services:

Most Productive Journals: There were total 967 journals covered by all the three indexing service viz ISA, CABI and INSPEC, during the whole study period (2005-2009). The first 50% papers (around 47611) were indexed in first two years. The top most 5 journals publishing Indian Papers were Environ Ecol followed by Asian J Chem , Acta Cienc Indica – Math and Curr Sci .

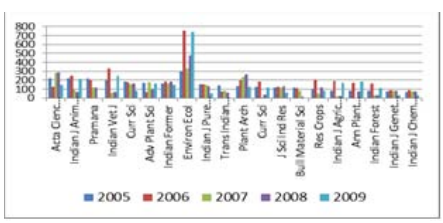


Figure 3. Top 20 Journals during 2005-09

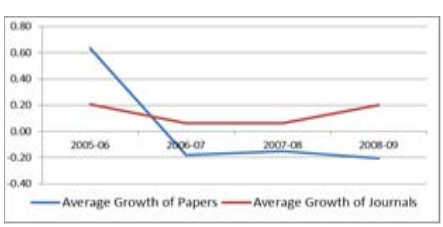


Figure 4. Average Growth of Indexed Journals having Indian Papers

The data indicates that the Average Growth of the Papers and the Indexed Journals by the services viz. ISA, CABI & INSPEC are adversely proportional to each other. The data has been analysed to compute the percentage of Indexed Journals by all the three Indexing Services individually also.

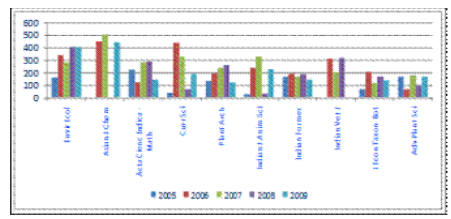


Figure 5. Top 10 Journals of ISA during 2005-09

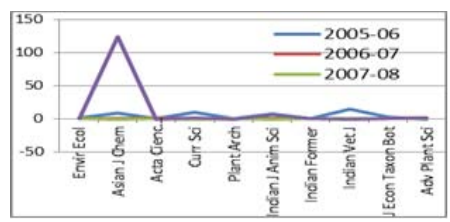


Figure 6. Average Growth Rate of Top 10 Journals of ISA during 2005-09

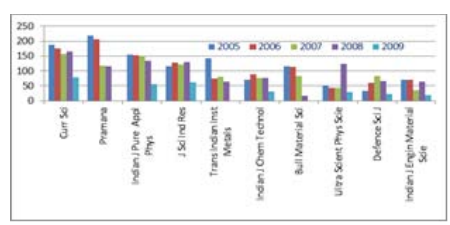


Figure 7. Top 10 Journals of INSPEC during 2005-09

There were total 72 Journals (publishing Indian papers) covered by INSPEC during 2005-09. The year wise coverage of the Journals was 50, 59, 54, 46 and 21 during the period of 2005-09 respectively.

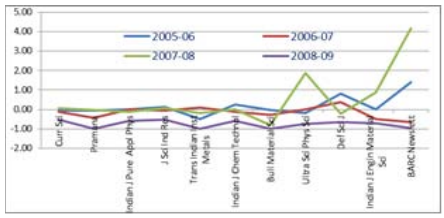


Figure 8. Average Growth Rate of Top 10 Journals of INSPEC during 2005-09

The graph of 'Average Growth Rate' of INSPEC, indicates that over the years , the coverage of journals having Indian papers have increased.

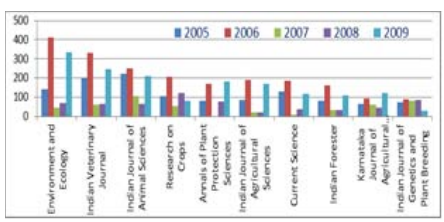


Figure 9. Top 10 Journals of CABI during 2005-09 According to the Publication

There were total 72 Journals (publishing Indian papers) covered by CABI during 2005-09.

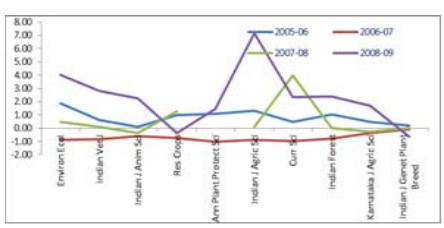


Figure 10. Average Growth Rate of Top 10 Journals of CABI during 2005-09

The trend of 'Average Growth' of Indian Papers and the Journals covered in CABI, indicates that during the period of 2006-2007, it has gone into a negative value, which calls for immediate attention of the authorities to look into the matter.

ISA being only Indian Indexing Service has covered maximum Indian Papers, whereas the coverage by other two Global Secondary Services indicates that over the years the coverage of the Indian Papers are not up to the mark.

Peer Review: The most important measure to ensure quality of a journal is the peer review system. It has been criticized that many Indian journals do not have peer review system. Therefore, in the current initiative, the peer review system followed by Indian journals is being investigated. To understand the peer review system, the following practice is followed.

Conclusion

We are concerned with poor peer reviewing practice of Indian journals. Unless the institutions insist their scientists to opt for publications in the peer reviewed journals and to consider the publications in the peer reviewed journals only, the Indian journals will continue to be in the vicious circle. Funding agencies should try to allocate funds in the less accessible geographical areas to encourage scientist of those particular Cities/States to come forward and carry out research activity followed by publications.

NANOSCIENCE AND NANOTECHNOLOGY IN SCOPUS: JOURNAL IDENTIFICATION AND VISUALIZATION

Teresa Munoz-Ecija, Benjamín Vargas-Quesada, Zaida Chinchilla-Rodríguez, Antonio J. Gómez-Nuñez and Félix de Moya-Anegón

[1] *teresamunyozecija@gmail.com*
SCImago Research Group Associated Unit, Granada (Spain)

[2] *benjamin@ugr.es*
University of Granada – Information & Communication Department (Spain)

[3] *zaida.chinchilla@cchs.csic.es*
CSIC – Institute of Public Goods and Policies (IPP), Madrid (Spain)

[4] *anxusgo@gmail.com*
SCImago Research Group Associated Unit, Granada (Spain)

[5] *felix.demoya@cchs.csic.es*
CSIC – Institute of Public Goods and Policies (IPP), Madrid (Spain)

Introduction

This document presents a new query for the retrieval of information about Nanoscience & Nanotechnology (N&N) based on the combination of previously published search strategies, contrasted by the scientific community. It led us to identify 80 core journals of N&N in Scopus, then map and analyze the underlying structure of N&N output using visualization techniques. N&N is established as a productive young discipline, crosscutting other fields in its rapid evolution, for which reason it has poorly defined limits to date, and needs some time to consolidate its identity as a discipline.

Materials & Methods

On 06/11/2012 we launched a search against the Scopus database using the combination of different queries proposed previously. Query 1 included terms with the root nano*, while excluding any terms containing the root nano* yet not related with N&N. To this end, we referred to the combination of proposals made by Noyons et al., 2003; Glänzel et al., 2003; Huang et al., 2003; and Meyer, Debackere and Glänzel, 2010**.** In turn, query 2 combined instruments and processes utilized in N&N with different types of materials, functions and other terms, in the wake of proposals by Noyons et al., 2003, Glänzel et al., 2003, Kostoff, Koytcheff and Lau, 2007 and Porter et al., 2008. Finally, query 3 included a series of terms related with N&N and was based on the work of Noyons et al., 2003; Glänzel et al., 2003; Huang et al., 2003; Kostoff, Koytcheff and Lau, 2006; Porter et al., 2008; and Lv et al., 2011. The combination of these three searches led to a new query which we believe can be perfectly adopted for the

identification of N&N documents in any specialized or multidisciplinary database (Annex). It is available too in Scopus format at: *http://www.scimago.es/benjamin/nanoquery.pdf*

Results

The total number of retrieved documents was 142,102: 70,726 articles, 30,314 conference papers and 4,062 reviews. The total number of references was 2,903,543. To identify the core journals, we selected those that reflected over 1% of total citation, eliminating multidisciplinary journals such as *Nature, Science* and *PNAS*. We aggregated all journals covered by Scopus that had the term "nano" in their titles and that had been cited at least once in 2010, with the understanding that the term "nano" in a title indicates that the journal pertains to the discipline of N&N and has been previously reviewed and validated by the scientific community (Schummer, 2004). The core journals of N&N identified amount to 80 (Table 1). Its visualization was achieved by means of Vosviewer (Van Eck & Waltman, 2010). It draws together the journals in four clusters (Figure 1). The correspondence between colors and journals is indicated in Table 1: red is 1, green is 2, blue is 3 and yellow is 4.

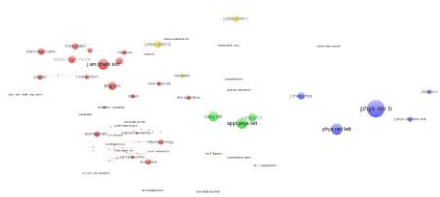


Figure 1. Scopus core journals in N&N

Discussion

The 80 journals indicated constitute the core group of N&N publications. However, for future reference, we ought to include in brackets the other 11 that also had the term "nano" in their title, since they will eventually be cited as well. Indeed, three of them appear as cited in the year 2010 in the SCImago Journal & Country Rank (Scimago, 2007), but not with reference to the documents we downloaded as the basis of our study. For a more complete view of core N&N journals, the listing and display offered here should be compared with a future contribution that also takes into account information from the Web of Science. One might expect a clearly outlined map of N&N journals in Scopus, given that core journals configure the basis of the map. Surprisingly, this is not the case. At first glance, there appears a fragmentation of journals revealing two major groups: Physics, condensed matter on the one hand, and Chemistry multidisciplinary on the other. This depiction implies that N&N is a highly transversal discipline, and that its borderlines are not well established.

Conclusions

N&N configure a highly transversal discipline, whose borders appear to defy delimitation. While most N&N documents are published in Physics and Chemistry journals, they may also be included in journals specializing in Materials Science or other slippery subject areas that may include the term "nano" in their title, and by virtue of this prefix, come to form the ranks of this discipline in high gear. Just 50 years old, it can be seen as a field in constant evolution, in parallel with the growth of journals who divulge its findings. It is a matter of time, and space, but N&N will eventually have its own profile in Scientometric mappings, That is, it will consolidate an identity as a distinctive scientific discipline, and be delimited as a separate category in databases such as

Scopus. Our analysis ventured into multidisciplinary databases by different means, recovering more documents than other attempts reported previously. To our knowledge, this is the first research study to combine all the approaches to N&N published to date, and discarding duplications. The sensitivity of this new type of query makes it adaptable to any database, as long as the syntax and operators are adjusted accordingly.

References

Glänzel, W., Meyer, M., Du Plessis, M., Thijs, B., Magerman, T., Schlemmer, B., Debackere, K., Veugelers, R. (2003) *Domain Study 'Nanotechnology: Analysis of an Emerging Domain*. Steunpunt O&O Statistieken.

Huang, Z., Chen, H., Yip, A., Ng, G., Guo, F., Chen, Z.K., Roco, M.C. (2003). Longitudinal patent analysis for nanoscale science and engineering: Country, institution and technology field. *Journal of Nanoparticle Research*, 5(3-4), 333-363.

Kostoff, R.N., Koytcheff, R.G, LAU, C.G.Y. (2007). Structure of the Global Nanoscience and Nanotechnology Research Literature. *Journal of Nanoparticle Research*, 9, 701-724.

Kostoff, R.N., Murday, J.S., Lau, C.G.Y., Tolles, W. M. (2006). The seminal literature of nanotechnology research. *Journal of Nanoparticle Research*, 8(2), 193-213.

Lv, P.H., Wang, G.F., Wan, Y., Liu, J., Liu, Q., Ma, F.C. (2011). Bibliometric trend analysis on global research. *Scientometrics*, 88(2),399-419.

Meyer, M., Debackere, K., Glänzel, W. (2010). Can applied science be 'good science'? Exploring the relationship between patent citations and citation impact in nanoscience.*Scientometrics*, 85(2), 527-539.

Noyons, E.C.M., Buter, R.K., Van Raan, A.F.J. (2003). *Mapping Excellence in Science and Technology across Europe: Nanoscience and Nanotechnology: Final Report*. European Commission. Retrieved March 2, 2012 from ftp://ftp.cordis.europa.eu/pub/nanotechnology/docs/ec_mapex_nano_final_report.pdf

Porter, A.L., Joutie, J., Shapira, P., Schoeneck, D.J. (2008). Refining search terms for nanotechnology. *Journal of Nanoparticle Research,* 10(5), 715-728.

Schummer, J. (2004). Multidisciplinarity, interdisciplinary and patterns of research collaboration in nanoscience and nanotech-nology. *Scientometrics*, 59(3), 425-465.

SCImago. (2007). SJR — SCImago Journal & Country Rank. Retrieved January 11, 2013 from http://www.scimagojr.com

Van Eck, N.J. and Waltman, L. (2010). Software survey: VOSviewer, a computer program for bibliometric mapping, *Scientometrics*, Vol. 84 No 2, pp. 523-538.

Annex

Query: ((nano* AND NOT (nano2 OR nano3 OR nano4 OR nano5 OR nanosecon* OR nano-secon* OR nanogram* OR nano-gram* OR nanomol* OR nanophtalm* OR nanomeli* OR nanogeterotroph* OR nanoplankton* OR nanokelvin* OR nanocurie OR nano-curie OR nanos OR nanos1 OR nanoproto* OR nanophyto* OR nanoflagel* OR wnanomol* OR wnano-curie* OR wnanocurie* OR anos1 OR nanobacter* OR nano-bacter* OR nanospray* OR nano-spray* OR plankton* OR n*plankton OR m*plankton OR b*plankton OR p*plankton OR z*plankton OR nanoalga* OR nanoprotist* OR nanofauna* OR nano*aryote* OR nanoheterotroph* OR "nanook of the north" OR nano-bible)) OR ((atomic-force-microscop* OR afm OR transmission-electron-microscop* OR tem OR scanning-tunneling-microscop* OR tunnel*-microscop* OR stm OR scanning-electron-microscop* OR sem OR self-assembl* OR selfassembl* OR self-organiz* OR edx OR energy-dispersive-x-ray OR energy-dispersive-x-ray-spectroscop* OR scanning-probe-microscop* OR electron-energy-loss-spectroscop* OR eels OR high-resolution-tem OR high-resolution-transmission-electron-microscop* OR uv-vis OR x-ray-photoelectron* OR x-ray-photoelectron* OR xps OR uv-visible-spectroscop* OR Ultraviolet-visible-spectroscop* OR hrtem OR Chemical-force-microscop* OR CFM OR scanning-force-microscop* OR SFM OR NSOM OR NEAR-FIELD-SCANNING-OPTICAL-MICROSCOP* OR SNOM OR "chemical vapor deposition" OR CVD OR " chemical vapour deposition " OR XRD OR " x-ray diffraction " OR " differential scanning calorimetry " OR DSC OR " molecular beam epitaxy " OR "mbe")) AND (surface* OR film* OR layer* OR substrate* OR roughness OR monolayer* OR mono-layer* OR molecul* OR structure* OR resolution OR etch* OR grow* OR silicon OR si OR silicium OR "silicon oxide" OR sio2 OR deposit* OR particle* OR formation OR tip OR atom* OR gold OR au OR polymer* OR copolymer* OR co-polymer* OR gaas OR inas OR superlattice* OR adsorption OR absorb* OR island* OR size OR powder* OR resolution OR quantum* OR multilayer* OR multi-layer* OR array* OR mater* OR supramolecular* OR biolog*)) OR (quantum-dot* OR quantum-wire* OR quantum-well* OR quantum-effect* OR "quantum computing" OR coulomb-blockade* OR coulomb-staircase* OR molecul*-motor* OR molecul*-ruler* OR molecul*-device* OR "molecular beacon" OR molecular-sensor* OR "molecular engineering" OR molecular-electronic* OR molecular-manufact* OR "molecular modeling" OR "molecular simulation" OR molecul*-wire* OR molecular-sieve* OR biosensor* OR bionano* OR hipco OR molecular-template* OR carbon-tub* OR carbontub* OR bucky-tub* OR buckytub* OR fulleren* OR biochip* OR dna-cmos* OR graphen* OR graphit* OR single-molecul* OR langmuir-blodgett OR pdms-stamp* OR pebbles OR nems OR quasicrystal* OR quasi-crystral* OR sol-gel* OR solgel* OR dendrimer* OR soft-lithograph*)). We limited the documents of interest to those published within the year 2010 and to articles, conference papers and reviews.

Table 1. Scopus core journals in N&N

Abbreviated journal title	*Cites*	*% Cites*	*Clusters*
phys rev b	184,370	13.879	3
phys rev lett	81,585	6.141	3
appl phys lett	75,445	5.68	2
j am chem soc	74,126	5.58	1
nano lett	60,151	4.528	2
Langmuir	42,982	3.236	1
j appl phys	42,849	3.226	2
j phys chem b	35,639	2.683	4
Angew chem int edit	35,435	2.67	1
adv mater	32,163	2.42	1
j chem phys	31,114	2.342	3
macromolecules	28,069	2.113	1
j phys chem c	27,568	2.075	4
Chem. Mater	24,481	1.84	1
biomaterials	21,191	1.6	1
anal chem..	19,287	1.45	1
j phys-condens mat	17,648	1.328	3
Chem. Commun	16,912	1.27	1
Polymer	16,692	1.257	1
Nanotechnology	16,346	1.23	1
Chem. Rev	15,008	1.13	1
nat mater	14,218	1.07	4
j mater chem.	13,431	1.011	1
physica e	12,658	1.008	3
Carbon	12,150	1.006	1
thin solid films	12,006	1.001	1
Chem. phys lett	11,836	1	1
nat nanotechnol	7,697	0.579	1
Acs nano	6,242	0.47	1
j nanosci nanotechno	3,799	0.286	1
j nanopart res	1,800	0.135	1
Nanomedicine	1,605	0.121	1
ieee t nanotech	968	0.073	1
nano today	820	0.062	1
Nanoscale res lett	700	0.053	1
nano res	643	0.048	1
microfluid nanofluid	564	0.042	2
Int j nanomed	548	0.041	1
j comput theor nanos	435	0.033	1
Nanomed-nanotechnol	352	0.026	1
j biomed nanotechnol	328	0.025	1
j. nanobiotechnology	299	0.023	1
Nano	293	0.022	1
curr nanosci	272	0.02	1
Nanoscale	266	0.02	1
Nanotoxicology	251	0.019	1
ieee t nanobiosci	232	0.017	1
Dig j nanomater bios	211	0.02	1
Int j nanotechnol	198	0.015	1
j nanomater	147	0.011	1
j nanophotonics	133	0.01	4
Int. j. nanosci.	127	0.01	2
Synth react inorg m	121	0.009	1
Wires nanomed nanobi	75	0.006	1
Micro nano lett	74	0.006	2
photonic nanostruct	71	0.005	4
nanosc and nanotech – asia	68	0.005	1
j exp nanosci	66	0.005	1
Fuller nanotub car n	63	0.005	1
j nanoelectron optoe	53	0.004	2
nanosc microsc therm	51	0.004	2
Nanoethics	48	0.004	1
j micro-nanolith mem	47	0.004	2
recent pat nanotech	36	0.003	1
Nami jishu yu jingmi gongcheng	33	0.002	1
j laser micro nanoen	30	0.002	2
j nano res	25	0.002	1
nano biomed. eng.	23	0.002	1
Int. j. nanomanufacturing	21	0.002	1
Nanotechnology law & business	20	0.002	1
Iet nanobiotechnol	18	0.001	1
e-j. surf. sci. nanotech.	16	0.001	1
ieee nanotechnol. mag.	16	0.001	1
Int. j. nanoparticles	15	0.001	2
nanotechnol. russ.	14	0.001	4
nanosci nanotech let	12	0.001	1
j. bionanosci.	11	0.001	2
j. nanostruct. polym. Nanocomp.	9	0.001	1
nanotechnol. sci. appl.	6	0.001	1
proc. Inst mech eng part nj n&n.	5	0.001	1

A NEW APPROACH FOR AUTOMATED AUTHOR DISCIPLINE CATEGORIZATION AND EVALUATION OF CROSS-DISCIPLINARY COLLABORATIONS FOR GRANT PROGRAMS

Ilya Ponomarev[1*], Pawel Sulima[1], Jodi Basner[1], Unni Jensen[1], Joshua Schnell[1], Karen Jo[2], Larry A. Nagahara[2], Jerry S.H. Lee[2], and Nicole M. Moore[2]

[*] *ilya.ponomarev@thomsonreuters.com*
[1]Thomson Reuters, 1455 Research Blvd, Rockville, MD (USA)

[2]Office of Physical Sciences – Oncology, Center for Strategic Scientific Initiatives, Office of the Director, National Cancer Institute, Bethesda Maryland, 20892 (USA)

Introduction

Integration of new knowledge often occurs at the interface of well-established research disciplines where cross-disciplinary thinking is rapidly becoming an integral feature of research. Many government programs fund research with the specific goal of bringing two fields together and the enhancement of cross-disciplinary research (CDR) collaboration. Evaluating the impact of these programs on collaborations and subsequent field convergence remains one of the least-understood aspects due to lack of indicators (Klein, 2008).

Development of a proper science classification scheme, suitable to a particular evaluation, is a necessary first step. There are several science classifications schemes in scientometrics, each targeted to different levels of research field aggregation. The most commonly used classification scheme is the Web of Science (WoS) subject category (SC) classification that consists of 265 SCs. In this scheme, a journal is assigned to one or more research areas. Publications are not directly assigned to research areas. Instead, the journal in which an article published determines the subject discipline(s) to which the publication belongs. Other classifications range from fine keyword/co-citation clustering into 500 plus disciplines (Boyak, 2009) to aggregation into broad research fields (22 in the Thomson Reuters Essential Science Indicators (ESI) or 13 used by the US National Science Foundation).

Challenges to measuring the impact of cross-disciplinary collaborations stem from an inability to detect convergent shifts within one standard science classification scheme. To spot these subtle changes one often needs to compare research categories at different levels of aggregation: from very broad (like “Life Sciences”) to very narrow (like “Remote Sensing”). Another problem is in the heterogeneity of subject categories themselves: some of them are narrowly focused while others, particularly those related to emerging fields, already have a more cross-disciplinary origin.

To address these problems and capture the uniqueness of CDR programs, we

developed an automated approach for assignment of scientific publications into disciplinary categories tailored for a specific grant program. In this case study, the approach was used to evaluate cross-disciplinary collaborations within the National Cancer Institute's (NCI) Physical Sciences – Oncology Centres (PS-OC) program. The program, established in September 2009, supports a network of twelve cross-disciplinary centres. The PS-OC network brings together over 150 investigators stemming from disparate fields of physics, mathematics, chemistry, engineering, cancer biology and clinical oncology with the primary goal of uniting the fields of physical science with cancer research to better understand the physical and chemical forces that shape and govern the emergence and behaviour of cancer at all levels.

Data and Methodology

Research advances, scientific outputs, and collaborative activities of PS-OC Centres are monitored through comprehensive semi-annual progress reports (60-80 pages). Data were collected from 166 active PS-OC investigators whose specialization was known prior to grant by self-nomination. 601 publications reported in the semi-annual progress reports (Fall 2009 - 2012) were verified and compiled for subsequent analysis. For a comparison, an additional list of 3,367 "baseline publications" was generated from WoS for PS-OC investigators prior to the PS-OC program (2006-2008). For both lists, more than 202,000 references were collected from 4,199 different journal titles.

Our goal was to automatically assign research field interest(s) to each investigator based on the research content of a person's publications during a certain period of time. Then, by monitoring new publications, we could trace how investigator research interests and CDR collaboration shifts with time.

We developed a three-step automatic assignment of investigators' scientific interests to *two or three* PS-OC program specific research fields (Physical Sciences, Oncology, and Life Sciences).

Stage 1: Mapping of SCs to 6 Intermediate fields

265 standard journal SCs were manually divided into six intermediate broad categories (IBCs) relating to the PS-OC cross-disciplinary research program (*physical sciences (PS), life sciences (LS), medicine (MED), oncology (OC), multidisciplinary (MU), and other (OTH)*). Such an intermediate classification is needed for proper weighting of publications from multidisciplinary journals and for additional renormalization of LS and PS disciplines.

Stage 2: Calculate IBC weights for each publication and aggregate weights at author level

IBC relevance weights for each publication were calculated based on the journal SC mapping of the publication itself and references in the publication. We assumed that if a paper published in journal A has *n* research fields, then the topic of this paper is equally distributed between these *n* IBCs. Thus each of these *n* fields will receive a weight equal to *1/n*, and the remaining fields are assigned a zero weight. For example, the journal "Radiation Research" is assigned three WOS categories ("Biology", "Biophysics", and "Radiology, NM and MI"). These three SCs were mapped to two distinct IBCs

(LS, PS). Therefore, the weights of the journal are *W(PS)=1/2, W(LS)=1/2, W(OC)=W(MED)=W(MU)=W(OTH)=0* . Such an assignment gives normalized weights. For each paper, journal weights and weights of references were first calculated separately, and then a combined weight was calculated. These weights were then aggregated for all author's papers and normalized. The same procedure is repeated for all references in an author's publications.

Stage 3: Ranking and category assignments

We ordered IBC weights in descending order. For the final assignment of a person's discipline (either the two or three classification scheme) weights of MED, MU, and OTH were proportionally redistributed into OC, PS and LS weights. Then the three (or two) weights were renormalized. For example, if a researcher had a weight distribution of OC=0.4, PS=0.2, LS=0.1, and MED=0.3 then after redistribution and renormalization the final weights were OC=0.57, PS=0.29, and LS=0.14.

Validation and Evaluation

At the beginning of the grant program all investigators identified themselves as an oncologist or physical scientist, and this information was used to validate the automated discipline assignment. Results were analysed using standard statistical precision/recall methods and are shown in Table 1. Both the precision and recall have very high values for the 2 discipline classification which validates the methodology. The precision and recall for the 3 discipline scheme have a lower but still acceptable level. It was identified that major discrepancies came from investigators who already had cross-disciplinary research interests before the PS-OC program started.

Table 1. Analysis of precision and recall for 3- and 2 disciplines schema.

	3 discipline classifications			2 discipline classifications	
Categories	**2006-2008 OC**	**2006-2008 LS**	**2006-2008-PS**	**2006-2008 (LS+OC)**	**2006-2008 (PS)**
NCI classificatior	**52**	**31**	**83**	**78**	**88**
Predicted researchers	29	83	54	76	90
Correctly predicted	24	26	52	71	83
Incorrectly predicted	5	57	2	5	7
Precision	0.83	0.31	0.96	0.92	0.93
Recall	0.46	0.84	0.63	0.94	0.91
F-Measure	**0.59**	**0.46**	**0.76**	**0.93**	**0.92**

Successful validation of our methodology helped us assess intra- and cross-disciplinary collaborations before and during program participation. Future directions aim to monitor changes in investigator research outputs after receiving funding.

Acknowledgments

We thank Julia DiCarlo and Yvette Seger for help with manuscript editing.

References

Boyack, K. W. (2009). Using detailed maps of science to identify potential collaborations. Scientometrics, 79(1), 27–44.

Klein, J.T. (2008), Evaluation of interdisciplinary and transdisciplinary research: a literature review. American Journal of Preventive Medicine, 35(2 Suppl), S116-23.

NORMALIZED INDICATORS OF THE INTERNATIONAL BRAZILIAN RESEARCH: A SCIENTOMETRIC STUDY OF THE PERIOD BETWEEN 1996 AND 2011

Maria Cláudia Cabrini Grácio[1] and Ely Francina Tannuri de Oliveira[2]

[1] *cabrini@marilia.unesp.br*
UNESP – Univ Estadual Paulista, 737 Hygino Muzzi Filho Avenue, 17525-900 Marília (Brazil)

[2] *etannuri@gmail.com*
UNESP – Univ Estadual Paulista, 737 Hygino Muzzi Filho Avenue, 17525-900 Marília (Brazil)

Introduction

The Brazilian science has grown comprehensively, particularly in the last 25 years (Glänzel, Leta & Thijs, 2006; Leite, Mugnaini & Leta, 2011). Within this context, to assess scientific production issues, the bibliometric studies constitute an objective approach, which offer a real comprehensive and true diagnosis about the scientific production of a specific area, of a group, of institutions or countries, producers of science and technology.

A country's scientific production analysis involves a broad group of bibliometric indicators which group themselves in production indicators, citation indicators and link indicators (Narin et al., 1994; Callon et al., 1993).

Despite the importance of the indicators for the analysis and the understanding of the contribution, insertion, and impact of a researcher or country for a knowledge area, a common problem in bibliometric analyses occurs when it's intended to hold comparative studies, given the specificities and peculiarities of each knowledge area. In this context, we focus on the relevance of the normalized indicators, which work as basis and enable comparative evaluations, either among areas or levels of aggregation, once they standardize the measurement units (Glänzel et al., 2009).

A normalized indicator may be defined as the quotient between the analyzed indicator, taken in its original value, divided by the indicator average in the studied scientific area (Moed, 2009). As result, this indicator standardizes the behavior of an individual, in a way that it situates it comparing it to the global tendency (average) observed in the area. In this research, individuals refer to countries. Value above 1 point that the individual (in this research, Brazil) presents a performance above what's expected for the group (in this research, countries).

In this research, we aim at comparatively analyzing the normalized scientometric indicators of production and citation of the Brazilian research, in the 27 areas of knowledge, from the data presented in the Scimago Journal & Country Rank gate, in the period between 1996 and 2011. Besides, we aim at grouping the areas by means of the similarities of the indicators analyzed.

Methodological procedures

The data were raised in the Scimago Journal & Country, by means of the following procedures: for each one of the 27 areas of knowledge, the total number of existing and Brazilian journals was taken in each one of them. We calculated the percentage of journals from Brazil compared to the existing total in the different areas. Next, for each area, the average of citable documents, citations per document and index h were calculated. Considering these values, for each area, the normalized Brazilian index was calculated for the indicators mentioned above.

Finally, we held a hierarchical cluster analysis, by Ward's method, in order to group the 27 areas analyzed, according to their similarities compared to the normalized indexes, considered together and simultaneously. The visualization of the clusters was achieved by means of the dendrogram.

Presentation and analysis of data

Table 1 shows that the performance of the Brazilian science is always above 1, indicating that Brazil has a position above the expected for the group of producing countries, for all the normalized indexes. We add that the normalized index h is the indicator with the smallest variation in the Brazilian performance.

We point out that, in average, the percentage of Brazilian indexed journals is 1.5% of the total of journals. Dentistry, Veterinary and "Agricultural and Biological Science" are the areas that present the greatest participation in Brazilian journals.

Regarding the normalized index of total of documents, the average of the areas is 3.2 times above the expected for producing countries. Veterinary, "Agricultural and Biological Sciences" and "Immunology and Microbiology" stand out as those which present scientific production much higher than the expected.

Table 1. Normalized indicators of production and of citation of the 27 areas related to Brazil.

Area	% of journals	***total documents***	***Cit per doc***	***index h***
Agricultural & Bio Scie	3.5	7.6	0.9	2.9
Arts and Humanities	0.8	2.1	1.1	2.0
Biochem.Gen&Mol Biol	0.8	2.9	1.0	2.5
Business. Manag &Acc	0.5	1.2	1.5	1.9
Chemical Engineering	1.2	2.8	1.2	2.7
Chemistry	1.1	2.8	1.1	2.6
Computer Science	0.2	2.1	0.9	2.6
Decision Sciences	0.6	2	1.3	2.3
Dentistry	5.7	2.3	1.4	3.9
Earth & Planet Sci	1.5	2.6	1.1	2.7
Econ. Econometr& Fin	1.2	1.5	1.9	2.0
Energy	0.0	2.3	0.9	2.5
Engineering	0.8	1.9	1.0	2.8
Environmental Sci	1.8	3.4	1.3	3.1
Health Professions	2.0	2.6	1.2	2.1
Immun & Microbiol	1.1	4.9	1.0	2.6
Materials Science	1.3	2.3	1.1	2.3
Mathematics	0.3	2.9	1.3	2.7
Medicine	1.3	3.9	0.9	2.8
Multidisciplinary	1.4	1.8	0.7	2.5
Neuroscience	1.6	3.5	0.4	2.9
Nursing	1.9	4.3	1.9	2.9
Pharmacology. Toxicology and Pharmaceutics	0.9	3.8	1.1	2.7
Physics and Astronomy	0.4	2.8	1.1	2.8
Psychology	2.5	2.3	1.3	2.5
Social Sciences	1.8	2.9	1.2	2.3
Veterinary	4.7	10.4	0.9	3.0
Mean	1.5	3.2	1.1	2.6

Source: authors' elaboration, from the ScimagoJr.

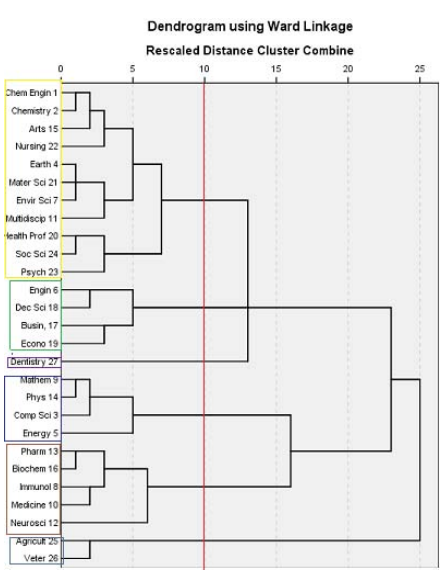
Concerning the normalized index of citation per document, it reached the average 1.1, value which is very close to the expected average (global) of the respective areas. Besides, this average is the lowest obtained among indicators analyzed in Brazil, aligning with the results found by Glänzel, Leta & Thijs (2006), which point to a visibility of the

Brazilian research lower than the average and an impact of citation relatively low.

The normalized index h presented average 2.6 times above the global average. We highlight the areas Dentistry, "Environmental Science" and Veterinary with the highest normalized indexes.

We present in Figure 1, the cluster of the 27 areas, grouped according to the similarities of the indicators of the Brazilian science, in which 6 groups are highlighted by color.

In relation to first group, the averages of every normalized indexes do not stand out in compared to the averages of the other 5 groups. Group 2, although presenting a less significant production, has an outstanding impact in compared to the other 5 groups. The group of Dentistry stands out from the others because it presents highest values of journals percentage indexed and index h. The group 4 presents the lowest average of percentage of journals indexed among constructed groups. The fifth group presents low index of citation per documents.



The sixth group presents the highest indexes of total of documents produced.

Final considerations

We point out that Brazil ranks above average in the group of producing countries, for all the normalized indexes under analysis. We observe that for the normalized index of total of documents, in all areas, Brazil presents its production above the global average of the respective areas.

References

CALLON, M., COURTIAL, J.P. & PENAN, H. (1995). *Cienciometría:* la medición de la atividad científica: de la bibliometría a la vigilancia tecnológica. Astúrias: Ediciones Trea.

GLÄNZEL, W., LETA, J. & THIJS, B. (2006). Science in Brazil. Part 1: A macro-level comparative study. *Scientometrics,* 67(1),67–86.

GLÄNZEL, W., et al. (2009). Subfield-specific normalized relative indicators and a new generation of relational charts: Methodological foundations illustrated on the assessment of institutional research performance. *Scientometrics,* 78 (1), 165–188.

LEITE, P., MUGNAINI, R. & LETA, J. (2011). A new indicator for international visibility: exploring Brazilian scientific community. *Scientometrics*, 88, 311 – 319.

MOED, H.F. (2009). New developments in the use of citation analysis in research evaluation. *Scientometrics*, 57, 13 –18.

NARIN, F., OLIVASTRO, D. & STEVENS, K. S. (1994). Bibliometric theory. practice and problem. *Evaluation Review*, 18, 65 – 76.

ON THE DEFINITION OF A REVIEW, AND DOES IT MATTER?

Robert Colebunders [1] and Ronald Rousseau[2]

[1] *bcoleb@itg.be*
Institute for Tropical Diseases, Antwerp, Belgium & University of Antwerp (UA), Belgium

[2] *ronald.rousseau@ua.ac.be*
University of Antwerp (UA), IBW, Venusstraat 35, Antwerp, Belgium

Introduction

In a previous paper (Colebunders et al., 2013) we investigated if the relative number of reviews in some medical fields is increasing and answered this question affirmatively, at least for the subfields Tropical Medicine, Infectious Diseases and Oncology. In that investigation we simply used the Thomson Reuters (WoS) definition of a review. According to http://thomsonreuters.com/products_services/science/free/essays/impact_factor/ any article containing more than 100 references is coded as a review. Articles in "review" sections of research or clinical journals are also coded as reviews, as are articles whose titles contain the word "review" or "overview." Yet, it is well known (Harzing, 2013) that Thomson Reuters' definition of a review is contested. In this contribution we consider two other definitions and check if using the other two definitions still leads to an increase in the relative numbers of reviews.

Methods

Data were collected during the first half of the month November 2012. Three definitions of a review publication were considered and results compared. The first definition is the Thomson Reuters or WoS definition of a review. The simplest alternative consists of retrieving all publications, classified by the WoS as an article or a review (eliminating in particular editorial material and meeting abstracts), that have either the word *review* or the word *overview* in the title. Finally, a third and broader alternative consists of all publications retrieved by a topic search (TS=) for the words review or reviews (we note that we did not use TS=review* as this would result in a number of false positives). The underlying idea of the third approach is that even if a review article does not have the word review in the title then it has sentences such as "This paper reviews … " or "We present a review of …" in its abstract. These three definitions were used to see if they lead to a significant difference.

For the determination of a possible increasing trend in the absolute and relative numbers of review publications we collected for each year over the period [1990, 2011] the total number of publications assigned to each of the three medical fields, and for each field and each year the number of reviews (according to the WoS definition), the number of (normal) articles (according to the WoS definition), the number of reviews according to our second definition and according to our third definition.

Results

Absolute and relative number of reviews and the definition of a review

We found an increase in the absolute number of publications in each of the three fields and a corresponding increase in the number of reviews, whatever the definition of a review. It is also clear that, in terms of publications, Oncology is the largest field and Tropical Medicine the smallest. When collecting the data we noticed the rather strange fact that a large number of publications (normal articles or reviews) that have the word review or overview in the title are not considered as reviews by the WoS, contrary to Thomson Reuters' website definition. Exact data are shown in Table 1. Surprisingly more than 40% are not considered to be reviews in the WoS. The fact that publications that are considered to be reviews by their authors are not considered as reviews by SCI and vice versa was already observed in (Aksnes, 2006).

Absolute and relative numbers of reviews differ depending on which of the three definitions are used. Most of the time definition 3 (topic search) yields the largest amount of publications while definition 2 (title words) yields the smallest. To save space we only show result for the field of Oncology (Table 2).

Increasing trends in the relative number of reviews

Now we come to the most interesting question, namely: is the relative number of reviews increasing over the latest twenty years? Considering the period [2000 – 2011] the percentage of reviews among all articles and reviews (WoS definition, not taking meeting abstracts, editorial material, and other items into consideration) is on average 10.2% in Oncology, 6.5% per year in Infectious Diseases and 3.6% in Tropical Medicine. These percentages are much higher than the 2.5% suggested by Price (1965). Yet, it is much more interesting to see if there really is an increasing trend over the period [1990, 2011]. As data are rather irregular we consider three-year moving averages (Fig.1).

Table 1. Publications having the word *review* or *overview* in the title. Period 1990-2011

Subject area	Classified as review	Classified as article	% reviews
Tropical Medicine	349	237	59.9
Infectious diseases	1533	1509	50.4
Oncology	4697	3468	57.5

Table 2. A: Number of publications classified as Oncology; B: Number of reviews Thomson Reuters definition); C: Number of reviews (def. 2: based on title); D: Number of reviews (def.3: based on topic search); E: Relative number of reviews (WoS definition); F: Relative number of reviews (based on title); G: Relative number of reviews (based on topic search).

Year	A	B	C	D	E	F	G
1990	13703	488	152	176	0.036	0.011	0.013
1991	14014	500	137	540	0.036	0.010	0.039
1992	14276	436	163	670	0.031	0.011	0.047
1993	16204	567	185	843	0.035	0.011	0.052
1994	16559	690	209	937	0.042	0.013	0.057
1995	18677	735	201	1019	0.039	0.011	0.055
1996	18520	872	240	1103	0.047	0.013	0.060
1997	21829	1070	323	1324	0.049	0.015	0.061
1998	24180	1204	339	1525	0.050	0.014	0.063
1999	25452	1209	294	1595	0.048	0.012	0.063
2000	22924	1413	280	1653	0.062	0.012	0.072
2001	24579	1542	299	1837	0.063	0.012	0.075
2002	27795	1648	290	1842	0.059	0.010	0.066
2003	28099	1904	342	2059	0.068	0.012	0.073
2004	39254	2036	399	2408	0.052	0.010	0.061
2005	44166	2266	456	2736	0.051	0.010	0.062
2006	44563	2702	495	3152	0.061	0.011	0.071
2007	46371	2791	577	3453	0.060	0.012	0.074
2008	51735	3085	599	3815	0.060	0.012	0.074
2009	52985	3206	666	4127	0.061	0.013	0.078
2010	50291	3322	728	4422	0.066	0.014	0.088
2011	49260	3433	791	4490	0.070	0.016	0.091

Table 3 shows the Pearson correlation coefficient and the slope of the regression line. According to the Thomson Reuters and the topic-based definitions there is always a clear increasing trend in the relative number of reviews. This trend is less clear or non-existing for the title-based definition.

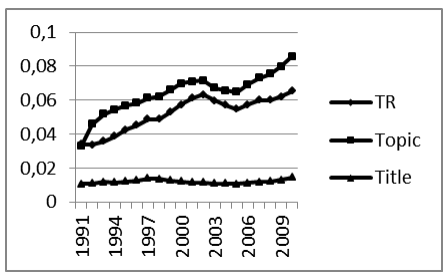


Figure 1. Percentages of reviews in the field of Oncology; TR: Thomson Reuters definition; Topic: results of Topic search; Title: results of Title word search

Discussion and conclusion

We have shown that, for the three medical fields we investigated, an increasing proportion of published scientific papers are review papers. This conclusion holds when using the Thomson Reuters definition of a review and for publications that are reviews according to a topic search. Remarkably the proportion of articles with the terms 'review' or 'overview' in the title shows little increase, at least in these three medical fields. As mentioned in (Colebunders et al., 2013) we consider this increase a disturbing trend and suggest that it is a consequence of the criteria used for evaluating scientists, departments and universities. To advance science we need more innovative research resulting in original research publications.

Table 3. Trends

Field	R (correlation coefficient)	Slope
Oncology		
WoS def.	0.91	0.0016
Title words	0.18	$3*10^{-5}$
Topic	0.90	0.0018
Infectious diseases		
WoS def.	0.89	0.0013
Title words	0.10	$3*10^{-5}$
Topic	0.97	0.0023
Tropical medicine		
WoS def.	0.88	0.0012
Title words	0.63	0.0003
Topic	0.78	0.0011

References

Aksnes, D.W. (2006). Citation rates and perceptions of scientific contribution. *Journal of the American Society for Information Science and Technology*, 57(2), 169-185.

Colebunders, R., Kenyon, C., & Rousseau, R. (2013). Increase in numbers and proportions of review articles in *Tropical Medicine, Infectious diseases* and *Oncology*. Submitted.

Harzing, A.-W. (2013). Document categories in the ISI Web of Knowledge: Misunderstanding the social sciences? *Scientometrics*, 94(1), 23-34.

Price, D.J. de Solla (1965). Networks of scientific papers. *Science*, 149, 510-515.

AN ONLINE SYSTEM FOR MANAGEMENT AND MONITORING OF EXTRAMURAL PROPOSALS FOR FUNDING BY ICMR – A CASE STUDY

Divya Srivastava[1], Vijai Kumar Srivastava[2], Aparna Govind Raj[3]

[1] *drdivya.srivastava@gmail.com,*
Division of Publications & Information, Indian Council of Medical Research, New Delhi 110029, India

[2] *vijaiksri@yahoo.com*
Division of Publications & Information, Indian Council of Medical Research, New Delhi 110029, India

[3] CDAC, Noida, India

Introduction

Management and monitoring of submitted proposals is a highly intellectual activity involving managerial and complex scientific activity. Submitted proposals often go through a process of scrutiny, getting reviewed at multiple levels by multiple people. Compiling these reviews and taking a decision on the proposal based on reviews is a very important process which requires skill and utmost care. Once accepted for funds, further monitoring & management of the research units and the individual researchers is an important task, which is basic to better decisions regarding future research management for policy decisions. The proposals are to be evaluated (both retrospective and prospective). This is reflected in the assessment criteria for past performance and future plans that reflect the main questions that need to be answered by the researchers / evaluators. Proposal submission, processing and management system helps perform all these tasks seamlessly and with ease. It helps better management of the entire process and aids better management of records. The system thus developed is unique for any Indian funding agency for processing and management of Research Proposals in the area of S&T including Medicine. Improvement, reducing the time lag, ease of approach, transparency and accountability are the main objectives of this system of quality assessment. Public accountability is both a requirement for publicly funded research and an inherent element in the improvement cycle in which this scheme of evaluation plays a dominant role. With regard to the objective of improvement, the system is directed toward both the research and its management. Evaluators are explicitly asked to judge not only the performance of an institute's research and researchers, but also its leadership, strategy and policy, and research organization. If applicable, the quality questions also may refer to the socio-economic impact of research and to multi- and interdisciplinary research. So far the system is running for last one year and a total of 1187 proposals have been received for funding. Out of these,

the evaluation committee has cleared a total of 902 proposals for further step. An analysis of the Major Discipline, Institute, City and the State indicates the trend of research activity in that particular geographical area. The data is being compared with the disease burden (categorized under different States) to infer that, if the research activity compliments the disease burden or not. As ICMR is the apex body for formulation, funding, coordination and management of Medical Research in India, this kind of conclusive results are very important for an 'Informed Decision Making' by the policy makers.

Background

There has been a tremendous growth in Information and Communication Technologies (ICT) in the last few years. ICT solutions provide better management and control to many business workflows thus enabling organizations to reap the benefits of more sophisticated tools that are used to perform their day to day operations The Indian Council of Medical Research (ICMR) also wanted to computerize their proposal processing and management workflow. ICMR is the apex body in India for formulation, coordination and promotion of biomedical research . One of the tasks it undertakes is to provide funds to organizations for carrying out research. For this purpose organizations across India submit their research proposals to ICMR. If ICMR approves of a research idea then it funds the corresponding organization to carry out the research work. Traditionally this entire workflow used to be performed manually. So ICMR was getting proposals as hard-copies, these are then posted to experts for their comments. The comments are received and a notification is sent back to authors. This approach has many drawbacks, a few listed below:

- Maintaining all received is a cumbersome task and the data is often prone to natural calamities and physical wear and tear. Also, the proposals require large amount of physical space for storage.
- Many a times proposals are lost or misplaced, thereby causing all the associated data to become completely inaccessible.
- Managing individual expert comments becomes very cumbersome.
- Searching for individual proposals is often a slow and cumbersome process.
- Performing any statistical analyses on the received proposals is an extremely difficult task, subject to a lot of errors since it is a manual process.

In order to improve efficiency of processing of its Extramural Research Program and to save efforts of the Investigators, ICMR has decided to adopt two stage processing of extramural projects. In this direction, ICMR has recently shifted from manual receipt and processing of extramural projects to on-line interactive system. The system has started functioning and all new projects *w.e.f.* 1.1.2012 are being received online. It may be mentioned that all further processing relating to reviews by experts, sanction and release of funds, report submission and final closure of the project would be through this on-line system only. All stake holders (PIs, experts, program officers, ICMR management disbursing officers) in the new system are interacting through this on-line system (available through the link in ICMR's website).

The system thus developed is unique for any Indian funding agency for processing and management of Research Proposals in the area of S&T including Medicine. Improvement, reducing the time lag, ease of approach, transparency and accountability are the main objectives of this system of quality assessment. Public accountability is both a requirement for publicly funded research and an inherent element in the improvement cycle in which this scheme of evaluation plays a dominant role. With regard to the objective of improvement, the system is directed toward both the research and its management.

Methodology

Every stakeholder is expected to log into the system after registering by providing some basic information about himself or herself. Furthermore, the access to senior officers of ICMR, Heads of Divisions, Program officers, experts and disbursing officers is being provided after proper registration into the system by everyone. Necessary assistance is being provided in this respect by ICMR Extramural Team. To make the system fully operational, the following procedure for ad-hoc extra-mural projects are being followed:

- The project proposals received in the AdHoc category is reviewed in two stages – first a concept proposal is asked from the Principal investigator and once this is approved a detailed proposal is asked for. Now in the new system, the concept proposal is submitted online. The detailed project is being accepted from the investigators whose first stage concept proposal has been accepted and conveyed to the investigator, again online.
- Each project submitted belongs to a Broad Area and a Major Discipline in the Broad Area. The list of Broad Areas is identified by ICMR based on the current trends in biomedical sciences.
- The identification of experts is done by Program officers. The Program Officers are members of ICMR who have expertise in specific major Disciplines in specific Broad Areas.
- The experts review the proposals and send back their comments. The comments are then reviewed and accordingly decisions are taken on each proposal. The decisions are then communicated to the respective PIs again on-line.

Observations and Salient Findings

The sophisticated value of online information provision is not to use the databases only for finding facts and accessing documents, but to tap the unique items of useful information, the nuggets of knowledge and (by synthesis and/or analysis) extract the "searched pattern" in the raw data. So far the system is running for last one year and the data has been extracted from the system, by generating need based reports on set parameters for analysis. In the present study an analysis of the in-house 'On-line System of Extramural Proposals' have been made to identify comparative upcoming subject areas as well as scientific hubs in India, in the field of Biomedical Sciences for the period of 2012 January- December 2012. The basic data for the study has been culled out from the system, related to the scope of the study only for 'ad-hoc' proposals being submitted from the different parts of India. The findings have clearly indicated change of productive institutions, subject areas being covered by investigators and the pattern of Cities and the 'Major

Discipline' being chosen by Investigators.

A total of 1316 proposals have been received for funding. An analysis of the Major Discipline, Institute, City and the State indicates the trend of research activity in that particular geographical area. The results include distribution of numbers of 'Investigators', demonstrate the presence of clustering in the 'networks', and highlight a number of apparent differences in the patterns of 'Selected Major Disciplines *vis-à-vis* affiliated Cities', Top most institutes in terms of share of proposals during the Year were: All India Institute of Medical Sciences (70), Post Graduate Institute of Medical Education and Research (62), CSM Medical University (40), Jawaharlal Institute of Post-Graduate Medical Education & Research (22), Indian Institute of Technology (20), Annamalai University (19), Manipal Univeristy,(18), Kasturba Medical College (16), Amity University (13), National Institute of Mental Health and Neurosciences (13), Panjab University (12), Narayana Medical College (11), National Institute of Nutrition (11), Maulana Azad Medical College (10) and Sanjay Gandhi Post Graduate Institute of Medical Sciences (10).

All the proposals were distributed in a total of 63 'Major Subject Disciplines'. These major disciplines were assigned on the basis of the 'Title' of the research proposals submitted by the individual investigators. Main subject areas covered by Investigators have also been analyzed. Percentage of papers published in top 10 major disciplines was 53.52 for the whole study period (Jan.-Dec. 2012). The top 15 major discipline of the submitted proposals were: Oncology, Medicinal Plant, Pharmacology, Nutrition, Microbiology, Endocrinology, Nanomedicine , Health System Research, Virus Disease, Bioinformatics, Neurological Science, Cardiovascular Disease, Social & Behavioral Research, Maternal Health and Cellular & Molecular Biology.

The Geographical Locations of Institutes were also analyzed. For the purpose the Cities were grouped under three main categories – Big Cities with full infrastructural Facilities, Medium Cities with developing stage and are up-coming with Institutions and Medical Colleges the last Category was of Smaller Cities, which lack proper infrastructure, funding, access to proper information etc.

During, the whole period, around 50% of the proposals were confined to Bigger Cities. The 10 topmost Cities were: Delhi (174), Chennai (93), Lucknow (88), Chandigarh (80), Kolkata (48), Bangaluru (42), Hyderabad (39), Manipal (36), Pune (34), Puducherry (33) and Mumbai with 29 proposals. Most of these Cities have contributed proposals in selected few areas.

The data is being compared with the disease burden to infer that, if the research activity compliments the disease burden or not. As ICMR is the apex body for formulation, funding, coordination and management of Medical Research in India, this kind of conclusive results are very important for an 'Informed Decision Making' by the policy makers.

The main findings from the analysis indicate

- There is a distinct relationship between institutional 'size and standing' and the number and distribution of major areas being studied by the investigators from that particular institute. On average, institutional analysis clearly indicated a strong invisible relationship with 'Major Disciplines Selected' and the size of the

institutions along with its geographical location.

- A greater proportion of proposals from smaller institutions than from larger institutions are concentrating in the 'Major Disciplines' associated with the 'Diseases & Health Issues' prevailing in those particular geographical locations.

Conclusion

The participation of the significant number of the scientists is evidence of a critical mass of researcher working on the related topics. The existence of the group has been considered as indicators of the maturity of research community. Due to the researchers' primary scientific orientation in their own discipline, their interests often are strongly related to their specialty. Researchers and information professionals in co-operation with domain specialists are often involved here in the studies of research frontiers; trends, gaps and similarities in research efforts at institutional, national, and international levels. Apart from this, these studies can be useful for science policy also. This is the smallest but financially the most potent group. Their studies are at the miso- and macro-level where the national, regional and institutional structures of science and their comparative presentation are in the foreground.

PAPERS PUBLISHED IN PNAS REFLECT THE HIERARCHY OF THE SCIENCES

Daniele Fanelli[1] and Wolfgang Glänzel[2]

[1] *dfanelli@staffmail.ed.ac.uk*
STIS-The University of Edinburgh, Old Surgeons' Hall, Edinburgh, EH1 1LZ, UK

[2] *Wolfgang.Glanzel@econ.kuleuven.be*
Centre for R&D Monitoring (ECOOM) and Dept. MSI, KU Leuven, Leuven, Belgium
Dept. Science Policy & Scientometrics, LHAS, Budapest, Hungary

Introduction

A long-standing controversy surrounds the idea that disciplines form a hierarchy in which, moving from physical to social sciences, studies become on average "softer" (e.g. Cole 1983, Simonton 2006). Several studies have proposed possible measures of scholarly consensus (e.g. Varga 2011) – a concept closely related to that of scientific "hardness".

We recently showed that many such consensus-related parameters support the hierarchy hypothesis, in a large random sample of papers published in 12 disciplines (Fanelli and Glänzel 2013, 2012). This latter study excluded multidisciplinary, high-impact journals like Nature, Science or Proceedings of the National Academy of Sciences (PNAS), so we wondered how papers in these latter journals would compare to the rest. In this poster we report preliminary results of a comparison with PNAS.

Materials and Methods

Analyses replicated exactly the procedure established in previous studies(Fanelli and Glänzel 2013, 2012), except for how papers were sampled.

Sample

We searched the PNAS archive for all papers classified by the journal in mathematics, astronomy, physics, chemistry, biochemistry, genetics, evolution, ecology, plant sciences, psychology, economic sciences, political sciences, anthropology and social sciences. We excluded papers that were classified in more than one category, and those that the Web of Science database classified as anything other than *Article*. Astronomy, physics and chemistry were grouped as physical sciences; genetics and evolution were considered "hard" biological sciences; the other biological disciplines as "soft" biological sciences; the remaining disciplines were classified as social sciences. The final sample consisted of N=2,008 papers.

Parameters

We measured the following parameters, identified as most relevant by previous analyses (Fanelli and Glänzel 2013, 2012):

1. **Number of authors** (log-transformed)
2. **Length of article** (total number of pages, log-transformed).
3. **Number of references** (total number of cited references, square-root transformed)
4. **Proportion of cited monographs** (books were identified by searching

each title and author in Google-Books).

5. **Price's index** (calculated on all references)
6. **Diversity of cited sources** (Shannon diversity of journal name, conference or book title)
7. **Relative title length** (total number of words in the title, divided by number of pages)
8. **First person use -singular** (proportion of singular personal pronouns – "I", "my", etc. – on the total words in the abstract)
9. **First person use -plural** (same as above, but with plural personal pronouns – "we", "our", etc.)
10. **Sharing of references - degree** (number of other papers with which at least one reference is shared, measured in a bibliographic coupling network).
11. **Sharing of references – average intensity** (average weight of links for each node, measured in the same bibliographic coupling network as above).

Analysis

The number of references shared between any two papers in the sample was counted by standard bibliographic coupling (Fanelli and Glänzel 2013)

The ability of each parameter to predict the hypothesised rank of a paper's discipline or domain was tested in a multiple ordinal regression model. In analogy with previous analyses, the number of references was excluded from main effects, to avoid collinearity, but was retained in the model as a weighting factor.

Results

Table 1 reports the main analysis on the PNAS sample, Table 2 reports results of the previous study, on specialized journals, and values obtained dividing the effect size estimates obtained of the two studies.

Table 1. Multiple ordinal regression, with domain rank as dependent variable. Bold highlights statistically significant effects (P<0.05) [Data sourced from Thomson Reuters Web of Knowledge]

PNAS (this study) N=2,008		
predictor	**b±se**	**z**
ln(n. authors)	0.039±0.027	1.420
Price's index	0.123±0.115	1.071
sqrt(Shannon, sources)	**1.963±0.163**	**12.031**
proportion of monographs	**4.443±0.224**	**19.835**
ln(1+n. pages)	-0.149±0.119	-1.253
ln(relative title length)	**0.191±0.053**	**3.612**
1st pers. singular	-23.31±14.89	-1.565
1st pers. plural	**-20.14±7.576**	**-2.659**
Single vs. multi-author dummy	-0.092±0.078	-1.178
log(1+sharing degree)	**0.228±0.018**	**12.786**
log(1+sharing intensity)	-0.055±0.052	-1.068
1st pers. singular *(sing vs. multi author)	23.39±14.87	1.573
1st pers. plural*(sing vs. multi author)	**18.70±7.57**	**2.471**

Effect estimates obtained on PNAS articles are remarkably similar to those obtained previously in specialized journals. Moreover, even though the sample size was much smaller, the effect of most predictors passed formal statistical significance thresholds (0.05). The Shannon-diversity of sources had a remarkably stronger effect in PNAS papers compared to other journals, but in most other cases effects measured in the PNAS sample were weaker, as shown by the ratio values in Table 2, which are mostly smaller than 1. The effect of four predictors had opposite sign in the PNAS sample: number of authors, Price's index, length of articles, and frequency of use of first person singular. These effects, however, were all relatively small and their confidence intervals overlapped with zero, so it is unclear whether such divergence reflects genuine differences between the two samples rather than just random

fluctuations. Overall, in any case, the direction and relative magnitude of effects measured in the two samples are very similar, as revealed by the high correlation of their z-scores (Pearson's r=0.796(95%CI: 0.436-0.936), t = 4.36, df = 11, P = 0.001).

Table 2 Multiple ordinal regression results from a previous study on specialised journals (Fanelli and Glänzel 2013), and ratio of regression estimates obtained in the PNAS sample and in the previous study. Bold highlights effects that have opposite sign in the two studies (i.e. negative ratio value). [Data sourced from Thomson Reuters Web of Knowledge]

Other journals (previous study) N=28,477			b(PNAS) / b(other)
Predictor	**b±se**	**z**	
ln(n. authors)	-0.088±0.01	-9.508	**-0.438**
Price's index	-0.069±0.03	-2.655	**-1.782**
sq.rt(Shannon, sources)	0.110±0.00	63.688	17.849
proportion monographs	7.505±0.05	165.223	0.592
ln(1+n. pages)	0.596±0.02	30.991	**-0.249**
ln(relative title length)	0.218±0.01	15.561	0.875
1st pers. singular	12.32±1.6	7.854	**-1.892**
1st pers. plural	-67.44±0.77	-87.429	0.299
Single vs. multi-author dummy	-0.303±0.01	-26.895	0.303
log(1+sharing degree)	0.252±0.00	70.329	0.904
log(1+sharing intensity)	-0.382±0.02	-21.786	0.145
1st pers. singular *(sing vs. multi author)	-17.74±1.57	-11.312	**-1.318**
1st pers. plural*(sing vs. multi author)	41.14±0.76	54.340	0.454

Conclusions

These results suggest that papers published in a high-ranking multidisciplinary journal like PNAS maintain most bibliometric properties hypothesised to reflect levels of consensus and/or "softness", although their values are shifted towards those of "harder" sciences.

Acknowledgments

Google-Books kindly increased its search limits to allow automatic searches. DF was funded by a Leverhulme Early-Career fellowship.

References

Cole, S. (1983), The hierarchy of the sciences? *American Journal of Sociology, 89*(1), 111-139.

Fanelli, D., & Glänzel, W. A Bibliometric test of the Hierarchy of the Sciences: Preliminary results. In In: Eric Archambault, Yves Gingras & Vincent Lariviere (eds.) Proceedings of the 17th International Conference on Science and Technology Indicators, Montréal, CA, 2012 (pp. 452-453): Science-Metrix and OST

Fanelli, D., & Glänzel, W. (2013). Bibliometric evidence for a Hierarchy of the Sciences. PLoS ONE, in press.

Simonton, D.K. (2006), Scientific status of disciplines, individuals, and ideas: Empirical analyses of the potential impact of theory. *Review of General Psychology, 10*(2), 98-112.

Varga, A. V. (2011). Measuring the semantic integrity of scientific fields: a method and a study of sociology, economics and biophysics. *Scientometrics, 88(*1), 163-177, doi:10.1007/s11192-011-0342-9.

A RESEARCH PROFILE FOR A PROMISING EMERGING INDUSTRY – NANO-ENABLED DRUG DELIVERY

Xiao Zhou,[1] Alan L. Porter,[2] Douglas K.R. Robinson[3] and Ying Guo[4]

[1]*belinda1214@126.com*
School of Management and Economics, Beijing Institute of Technology, Beijing (China)

[2] *alan.porter@isye.gatech.edu*
School of Public Policy, Georgia Institute of Technology, Atlanta (USA), and
Search Technology, Inc., Norcross GA (USA)

[3] *douglas.robinson@teqnode.com*
teQnode Limited, Paris (France), and Université Paris-Est, LATTS-IFRIS, France

[4]*violet7376@gmail.com*
School of Management and Economics, Beijing Institute of Technology, Beijing (China)

Background

With the global emphasis on the development of nanotechnology ("nano"), Nano-Enabled Drug Delivery ("NEDD") systems are rapidly emerging as a key nano application area. NEDD offers promise in addressing pharmaceutical industry challenges concerning solubility, cost-reduction, disease & organ targeting, and patent lifecycle extension. A combination of factors promotes nanoparticle-enhanced and other nano-facilitated drug and gene delivery systems.

Approach & Research Questions

Publications and patents can provide different (and complementary) insights for the same field of interest. We devise a multi-component search strategy to construct an NEDD dataset from the Web of Science (WOS), Medline, and the Derwent Innovation Index (DII).

We also attempt to address other research issues, like generating an inductive approach to figure out the subsystems; identifying the linkage among countries or top organizations; seeking the hot topics and estimate their future prospects. The details can be seen in figure 1.

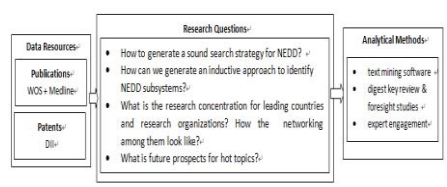


Figure 1. Approach & Research questions

We advanced a conceptual framework to approach NEDD, informed by various reviews and "foresight" pieces. This led us toward categorization to frame our current NEDD search (Table 1).

To retrieve a representative set of NEDD research records, the terms must be used in combination. We balance

retrieval (i.e., capturing a high percentage of the relevant records) with precision (i.e., without undue noise).

After considerable probing and consultation, we key on two general search strategies:

(1) D + P + N; (2) D + T + N.

We apply these terms in Web of Science (WOS), Medline, and the Derwent Innovation Index (DII) . The total results can be seen in Table 2.

Table 1. Nano-Enabled Drug Delivery: Related Terms

No.	Category	Keywords
1	B (biological processes)	(bioavailab* or biodistrib* or biocompatib* or cytotox* or biodegradab*)
2	I (imaging)	Image* or imaging
3	T (target)	(Cancer or tumor* or tumour* or "RNA interference" or RNAi)
4	H (helpers)	("polyethylene glycol" or pegylate or PEG or molecule* or polymer* or polyethylenimine or PEI or polyspermine or polypropylenimine or "poly lactic-co-glycolic acid" or PLGA or cyclodextrin or dendrimer* or chitosan* or atelocollagen* or "hyaluronic acid" or polypeptid* or peptid* or lipid* or ligand* or or Micelle* or Liposom* or conjugat* or Viral* OR Virus* or nonvira* or non-vira*)
5	P(pharmaceutical)	(1) (agent* or Drug* or pharmac* or formulation*) (2) (siRNA or "short interfering RNA") (3) microRNA (4) DNA or gene (5) (Dox or Doxorubicin*) (6) actives or adjuvant
6	D(delivery approach)	(1) (deliver* or vehicle* or carrier* or vector*) (2) (treat* or therap*) (3)("control* releas*" or transduct* or transfect* or transport* or translocat*)
7	N(nano-delivery vehicle)	This category means GT nano Database or some approximation of its search terms; also consider viral or virus or dendrimer or colloid

Table 2. Total NEDD dataset

Data type	Database	Number of records	Total number for different data type
Publication data	WOS	61,465	85311(54.6% overlap between WOS and Medline)
	Medline	52,529	
Patent data	DII	8426	8426

Subsystems

We combine content in title, abstract, keywords (authors) and Keywords Plus fields and consolidate term and phrase variations. Drawing upon co-occurrence, we use VantagePoint's (www.theVantagePoint.com) Principal Components Analysis (PCA) routine to group 585 frequently occurring and interesting terms. This factoring (PCA) yielded 19 topical groups of related terms. Colleagues knowledgeable about NEDD helped tune these to 21 major topics.We separate four major subsystems for NEDD: drug, nanocarrier, delivery outcome, and disease based on literature study (reviews) and text mining results. The we put 21 topics into four subsystems. Figure 3 shows four subsystems for NEDD.

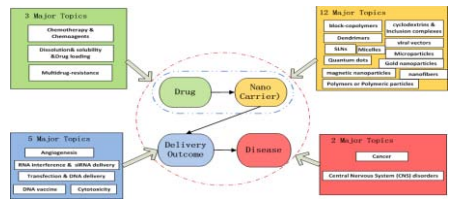


Figure 3. NEDD Subsystems

Figure 4 shows the trends for four subsystems.

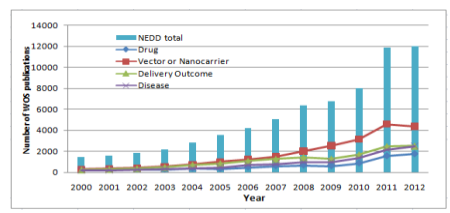


Figure 4. Trends for each NEDD Subsystem (WOS)

Top Organizations and Leading Countries Analysis

Figure 5 shows similarities among the 20 top research organizations based on relative emphases on 585 key terms. Ten of the top 20 organizations are in the US, with 6 in China. All six Chinese research organizations form one big group. CAS is the leader. This group focus on micelles and bio-copolymers. US research organizations can be distinguished into two small groups. They focus gene transfer, DNA and viral vector.

In addition, we choose three representative countries or region--China, US and Europe Union(EU) to compare. The total publication trends for them look similar. Figure 6 shows the trends

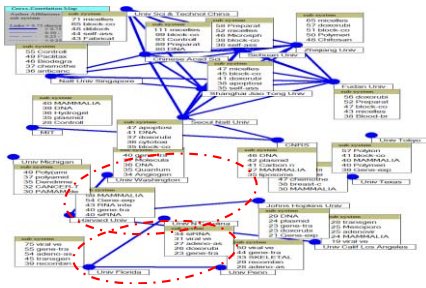


Figure 5. Cross-relationships among Organizations

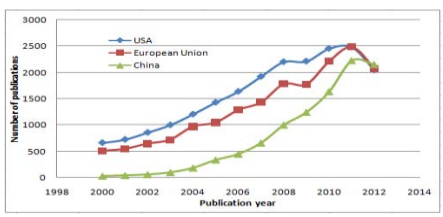


Figure 6. The total SCI trends for China, US and EU

Hot topics and their future prospects

There are four hot topics in recent years: RNAi, Cytotoxicity, magnetic nanoparticles, and gold nanoparticles. Figure 7 shows the trends. RNAi research has increased sharply since 2002. US, China, and Japan lead in this area. Cytotoxicity, magnetic nanoparticles, and gold nanoparticles increase sharply from 2007.

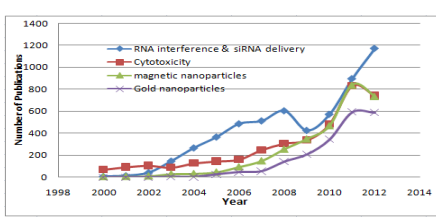


Figure 7. The Developmental Trend for Hot Topics

Acknowledgements

This research draws on support from the National Science Foundation (NSF) Science of Science Policy Program – "Revealing Innovation Pathways" (Award No. 1064146) to Georgia Tech, and also NSF support through the Center for Nanotechnology in Society (Arizona State University; Award No. 0531194). The findings and observations contained in this paper are those of the authors and do not necessarily reflect the views of the National Science Foundation.

Reference

Zhou, X., Porter, A.L., Robinson, D.K.R., and Guo, Y. (2013), Analyzing Research Publication Patterns to Gauge Future Innovation Pathways for Nano-Enabled Drug Delivery, *Portland International Conference on Management and Engineering Technology (PICMET), San Jose, California, 2013*.

Guo,Y., Zhou, X., Porter, A.L., Robinson, D.K.R. (2013), A comparative analysis of China vs. US: Two important players in the Nano-enhanced drug Drug delivery Delivery (NEDD) Race. *Portland International Conference on Management of Engineering and Technology (PICMET), San Jose, CA*.

THE P-INDEX: HIRSCH INDEX OF INDIVIDUAL PUBLICATIONS

István Papp[1], Mária Ercsey-Ravasz[2], Dávid Deritei[3], Róbert Sumi[4], Ferenc Járai-Szabó[5], Răzvan V. Florian[6], Alexandru I. Căbuz[7], Zsolt I. Lázár[8]

[1] *steve.prst@gmail.com,* [2] *ercsey.ravasz@phys.ubbcluj.ro,* [3] *deriteidavid@yahoo.com,* [4] *robert.sumi@phys.ubbcluj.ro,* [5] *ferenc.jarai@phys.ubbcluj.ro,* [8] *zsolt.lazar@phys.ubbcluj.ro*
Faculty of Physics, Babeş-Bolyai University, str. M. Kogălniceanu nr. 1, 400084 Cluj-Napoca, Romania

[6] *florian@epistemio.com,* [7] *cabuz@epistemio.com*
Epistemio Systems SRL, str. Cireşilor nr. 29, 400487 Cluj-Napoca, Romania
Epistemio LTD, 145-157 St. John Street, London EC1V 4PW, UK

Introduction

Evaluation of science is typically performed through peer review by experts in the field. However, evaluation is also performed through the use of bibliometric indicators, as a complement to peer review or as a replacement when there is a lack of resources to perform a proper peer review. The journal impact factor, eigenfactor (Bergstrom, West & Wiseman 2008, Davis 2008), article influence score (Bollen, Rodriguez & Van de Sompel 2006) or h-index (Hirsch 2005) are frequently utilized. One of the major shortcomings of these indicators is that they are not applicable across different disciplines (Seglen 1997, Bollen *et al.* 2009, Davis 2008, Fersht 2009).

For characterizing individual articles the most used indicator is simply the number of citations. Both the average number of references per article and the average time needed for an article to be cited differ widely between disciplines. This can cause extreme differences in the number of citations received by articles in different fields, hampering the use of this indicator for evaluations across domains. Also, the raw number of citations does not reflect the quality of these citations.

Here we show that a measure similar to the h-index could be used for the evaluation of individual publications. We used a citation network extracted from a large database of bibliographic information, including over 11 million scientific publications. The nodes of the network represent publications and directed links correspond to citations. The publication-level h-index, which we call the p-index, correlates with the logarithm of the number of citations, thus reducing the large differences between domains. We show in seven scientific fields that the distribution of the p-index features relatively small differences between the domains.

Results

Article-level h-index

The original Hirsch index (h-index) has been defined for the evaluation of scientists (Hirsch 2005) or scientific groups, such as departments. The h-index of a scientist is the maximal number *h* such that he/she has at least *h*

publications, which have at least *h* citations each. On Fig.1 we show how we adapted this definition to the case of individual publications. The p-index of a publication A is the maximal number *p* such that the publication is cited by at least *p* publications (B,C,D), which have at least *p* citations each. The central node A on Fig. 1 has p-index=3.

It is important that when counting the citations of B,C,D,E,F we do not include citations coming from nodes which already cite the main paper A. The link C→B on Fig.1 is not considered when counting the citations of B. When a small group of authors cite each other very frequently, the number of citations of papers can be large, however the p-index will not increase significantly unless their influence goes beyond their small inner circle.

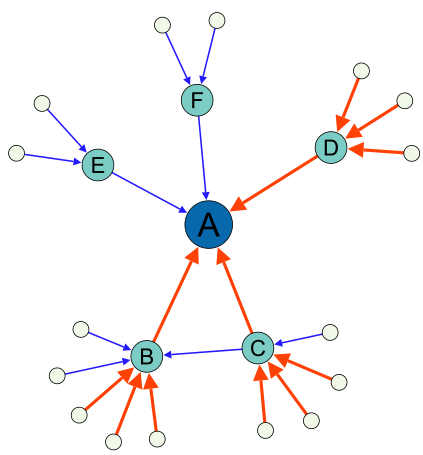


Figure 1. Article A has p-index 3 because it is cited by 3 articles (B,C,D) which themselves have at least 3 citations each.

Database

Our database was provided by Epistemio (www.epistemio.com) and includes scientific publications as well as a subset of their references. This subset is determined by the current availability of data (11,010,882 papers and 37,616,131 citations at the moment) and is unbiased with regard to scientific domain or publication date. The identified links (citations) give us a part of the real network, which can be considered as a random sampling and can already give us information about some statistical properties of the actual citation network. It is not easy to identify scientific domains to which the papers belong, because many journals are publishing papers from several different domains. However, based on the names of journals we identified seven smaller subsets of nodes (in total around 500,000 papers), which can be clearly associated with seven scientific domains: Physics (164,763 papers), Chemistry (40,173 papers), Engineering (38,261), Biology (109,158 papers), Medical Sciences (96,013) Mathematics (30,014) and Computer Science (2,241). On Fig. 2 we plot the distribution of the p-index of papers in these seven domains. The probability distribution shows an exponential decay and the differences between distributions for different domains are relatively small.

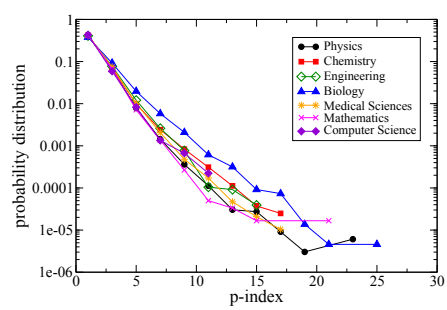


Figure 2. Probability distribution of the p-index of articles in the 7 different subgraphs associated with scientific domains.

Relation between p-index and number of citations

The most used indicator for evaluation of articles is the number of citations they received. In our graph this is given by

the in-degree (k_{in}) of the node: the number of incoming links. On Fig. 3 the colormap shows the correlation between the p-index and number of citations. From the definition itself it follows that a large p-index cannot be achieved with a small number of citations: on Fig. 3 we notice a line below which the number of articles is zero. The reverse, however, is not true: there are many articles having a lot of citations but a low p-index. This might indicate that articles farther from that separation line have citation number with less relevance to their actual impact.

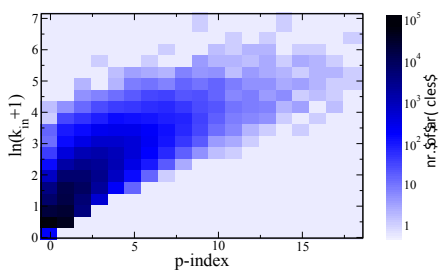


Figure 3. The colormap indicates the number of articles (on log-scale) with a given p-index and in-degree, k_{in}. Because k_{in} changes over several orders of magnitude, the y-axes shows its logarithm.

Conclusion

We have shown how the h-index can be adapted to evaluate individual publications. We used a citation network extracted from a large database to compute the p-index, an indicator that reflects not only the number of citations received by a paper but also their quality. While our database needs to be further improved the statistical results are already promising, indicating that the p-index could reduce the differences in the evaluation of different scientific fields, while offering a better estimation of the publications' real impact.

Acknowledgments

This work was supported by a grant of the Romanian National Authority for Scientific Research, CNDI–UEFISCDI, project number PN-II-PT-PCCA-2011-3.2-0895.

References

Bergstrom, C.T., West, J.D. & Wiseman, M.A. (2008). The eigenfactor metrics. *J. of Neuroscience,* 28, 11433-11434.

Bollen, J., Rodriguez, M.A. & Van de Sompel, H. (2006). Journal status. *Scientometrics*, 69, 669-687.

Bollen, J., Van de Sompel, H., Hagberg A. & Chute, R. (2009) A Principal Component Analysis of 39 Scientific Impact Measures. *PLoS ONE,* 4, e6022.

Davis, P.M. (2008). Eigenfactor: Does the Principle of Repeated Improvement Results in Better Estimates than Raw Citation Counts. *JASIST,* 59, 2186-2188.

Fersht, A. (2009) The most influential journals: impact factor and eigenfactor. *PNAS,* 106, 6883-6884.

Hirsch, J.E. (2005). An index to quantify an inidividual's scientific research output. *PNAS*, 102, 16569-16572

Seglen, P.O. (1997). Why the impact factor of journals should not be used for evaluating research. *BMJ,* 314, 498-502.

PRELIMINARY ANALYSIS OF THE FINANCIAL ASSISTANCE TO NON-ICMR BIOMEDICAL SCIENTISTS BY INDIAN COUNCIL OF MEDICAL RESEARCH (ICMR)

Sandhya Diwakar[1] and Keshari K. Singh[2]

[1] *sandhyadiwakar@gmail.com,* [2] *singhkeshari@yahoo.com*
Indian Council of Medical Research, Ansari Nagar, New Delhi - 110029 (India)

Introduction

Health research is the key to a well-functioning and effective health sector in the country. Major scientific breakthroughs hold the promise for more effective prevention, management and treatment for an array of critical health problems. The research to be undertaken should be on country specific health problems essential for the formulation of sound policies and plans for field action. Medical research in the country needs to be focused on new therapeutic drugs/vaccines for tropical diseases, normally neglected by multinational pharmaceutical companies on account of their limited profitability potential. In the Government sector, such research has been confined to the research institutions under the Indian Council of Medical Research, and other institutions funded by the Central/ State Governments.

Since its establishment, the ICMR has been making concerted efforts to address the health needs of the nation. The Council has discharged its national obligations through its network of 31 national institutes including Six regional medical research centres, over 100 field stations and a strong and vibrant extramural research in medical colleges and other institutes.

To provide an opportunity to academic scientists and trainees and to provide a stimulus for those working or contemplating working in the field of medical science, ICMR initiated an International Travel Grants Program in 2009 for Indian scientists to participate in international conferences, seminars, workshops and symposiums. The applicants should be Bio-medical scientist engaged in R&D work. Senior Scientists (above 35 yrs of age) working in academic institutions and research laboratories and young scientists (below 35 yrs of age) including medical graduates, post-graduates and research scholars are eligible to apply to international scientific events.

The budget sanctioned for the program was INR 3.0 Crores (about USD 550K), out of which the amount disbursed was INR 2.60 Crores (about USD 500K). During that period 1505 travel grant applications were received out of which 771 applications were approved for funding and 420 applicants finally availed the grant. Travel grants went to individuals at many institutions in the country and provided support for a wide range of biomedical research activities. An outcomes survey can be conducted to enhance the overall value and utility of the travel grant program.

Methodology

Data for the three year period during 2009-2012 was collected. Data points included name of the scientist, institution, designation, age, gender, state, conference title, venue, area of medical science, amount sanctioned/released and whether the application was approved, availed or rejected. The collated data was studied to identify the distribution of applications by state, area of medicine, designation; institution etc and inferences have been drawn from the study.

Observations

Zone-wise Status

A zone-wise analysis indicates that North Zone is the most active with highest number of applications for grants received, approved and availed.

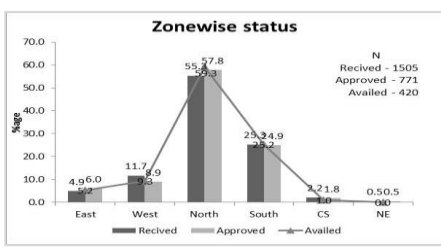


Figure 1. Zone wise distribution of applications received, approved, availed

New Delhi led all other states in terms of the applications submitted with a maximum number of 422. Karnataka ranked second with the 191 applications submitted followed by Uttar Pradesh (182), Tamil Nadu (120), Maharashtra (113) and Union territory of Chandigarh (106). These states accounted for three-fourth of the total applications received by ICMR.

Leading Research Institutions

Presence of top institutions in the states may explain why the some states have led other states. New Delhi is home to All India Institute of Medical Sciences (AIIMS), Maulana Azad Medical College, Jawaharlal Nehru University. National Institute of Mental Health and Neurosciences (NIMHANS) and Indian Institute of Science (IISc) are based in Bangalore in the southern state of Karnataka while Banaras Hindu University (BHU) is in the northern state of Uttar Pradesh. These institutions are amongst the top-tier research institutions of India. Further, out of the applicants who availed grants, those from AIIMS had a share of 14.8% followed by PGIMER, Chandigarh and NIMHANS, Bangalore at 7.1% each.

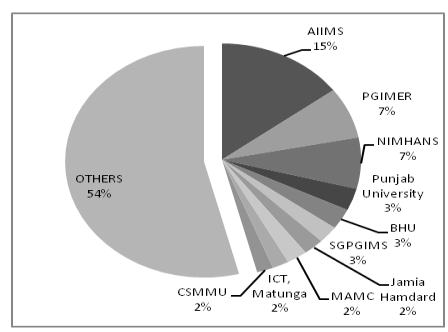


Figure 2. Top 10 institutes which availed grants, Total availed grants = 420

Research Areas

In terms of the bio-medical science discipline, it was noted that the applications were received in wide range of areas such as pneumonia, molecular oncology, cardiology, asthma, AIDs and so on. There were 206 unique areas under which scientists had submitted applications of which Oncology emerged as the top-most area for which this scheme was availed followed by Drug Development and Pharmacology.

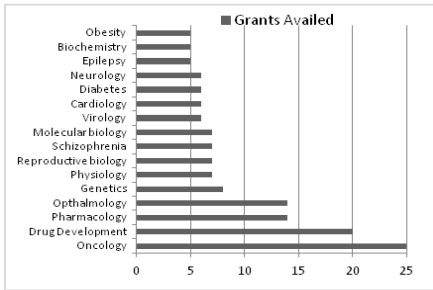


Figure 3. Top 16 research areas of availed scheme grants

Researchers Profile

Designation-wise analysis of the applicants who availed travel grants shows that the Research Fellows (JRF-SRF) had the highest share of 40.2%. The other major section of researchers benefitted from the scheme was that of Assistant Professors, which accounted for 10.3% of the availed grants.

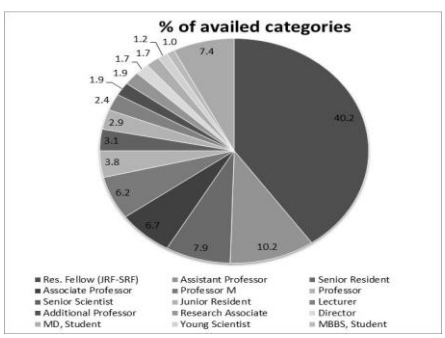


Figure 4. Designation-wise breakup of availed applications

Summary

Some general insights based on the sample data during the period 2009-2012 shows that out of the 420 selected proposals:

- New Delhi led all other states in terms of the applications submitted (422) and grants availed (133) followed by Karnataka with 191 applications submitted and 60 availed.
- AIIMS led the institutions availing grants and had a share of 14.8% followed by PGIMER, Chandigarh and NIMHANS, Bangalore at 7.1% each.
- Oncology emerged as the top-most area for which this scheme was availed followed by Drug Development and Pharmacology
- The Research Fellows (JRF-SRF) had the highest share of 40.3% followed by Assistant Professors who accounted for 10.3% of the availed grants.

Future directions

- Comparison of ICMR's funding program with similar programmes by other agencies domestically and internationally.
- Study Indian disease burden vis-a-vis financial support given to researchers for health research and collaboration.

References

Report of the Working Group on Health Research for XII Five Year Plan (September 2011). Department of Health Research and Family Welfare, 1-2.

Burroughs Wellcome Fund, www.bwfund.org

National Commission on Macroeconomic and Health

Health Information of India 2004,Government of India, Central Bureau of Health Intelligence, DGHS, MOH & FW, Nirman Bhavan, New Delhi-11 website: www.cbhidghs.nic.in

THE PRODUCTIVITY AND IMPACT OF ASTRONOMICAL TELESCOPES – A BIBLIOMETRIC STUDY FOR 2007 – 2011

Dennis R. Crabtree

Dennis.Crabtree@nrc-cnrc.gc.ca
National Research Council Canada, 5071 West Saanich Road, Victoria, BC (Canada)

Introduction

Since the telescope was invented in the early 17th century, astronomers have relied on increasingly complex and expensive instruments to further their studies of the Universe. The next generation of telescopes, with apertures of approximately 30-m, will cost more than $1B to construct. The main *product* of modern observatories are the publications in refereed journals based on data obtained using their telescope(s).

Increasingly, bibliometric techniques are being applied to the refereed papers produced by modern observatories. Observatory directors studiously compare their telescopes' performance with that of similar telescopes and the funding agencies anxiously wait for the return on their massive investment in these expensive facilities.

In this paper I will examine and compare the productivity and impact of eighteen telescopes using the basic bibliometric tools of paper and citation counts.

Input Data

Observatories carefully track the refereed papers that utilize data from their telescopes. Most observatories publish their list of *observatory publications* on the Web. The input data for this study is comprised of the list of observatory publications for nineteen of the largest optical/sub-mm telescopes used for astronomical research. The lists of papers published between 2007 and 2011 were gathered from the Web in most cases, but for some telescopes the lists were sent to me by observatory staff. These lists were incorporated into a custom designed Microsoft Access database/

Bibliometric Data

The international astronomy community is fortunate to have access to the NASA Astrophysics Data System (NASA ADS) (Kurtz, et al. 2000). The ADS provides bibliometric information that is used by all professional astronomers.

The NASA ADS database includes full publication information for each article (title, authors, journal, volume, page and year), as well as current citation counts. Each article in the system is assigned a unique bibliometric identifier (bibcode). This bibcode can be used to extract all the relevant information on that article from the ADS database.

For the work described in this paper, the correct bibcode was generated for each observatory publication and then the NASA ADS was queried to extract the publication information and the number of citations for that paper.

Productivity

The productivity of a telescope is the number of refereed papers published during a certain time period. Figure 1 shows the total publications per telescope for the 2007-2011 period. One can see that the productivity varies significantly between the telescopes. The main reason for the very low productivity of the LBT is that it only recently began operations and it takes up to 10 years of operation for a telescope to achieve full productivity (Crabtree and Bryson 2001)

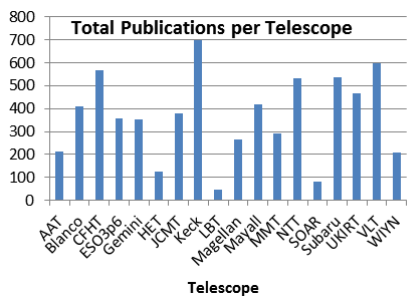


Figure 1. The total number of refereed papers published per telescope for the period 2007-2011

Impact

Citation counts are the most frequently used metric for measuring the impact, or relevance, of a refereed publication. A publication gathers citations over time so one can't really compare the raw citation counts of a paper published in 2007 with one published in 2011.

One approach to addressing this problem is to normalize the raw citation counts by a standard measure that increases with time similar to raw citation counts. I have used the citation count of the median paper published in the *Astronomical Journal (AJ)* as a standard measuring stick to normalize the raw citation counts. If there are 301 papers published in a given year, then the citation count of the 150th paper (ranked in descending citation count order) is the normalization factor **all** papers published in that year, regardless of the journal in which they are published.

I define the *impact* of a paper to be the number of citations to that paper divided by the citation count of the median *AJ* paper as defined above. This approach is very successful and allows papers of different publication years to be compared, and to compute aggregate impact metrics of papers published over a range of years.

One important measure of performance of a telescope's publications is the *average impact per paper (AIPP)*. Since the impact distribution of a telescope's paper is very non-normal (very long high-impact tail), the *median impact per paper (MIPP)* is also of interest. The difference of these two metrics between the telescopes is a good measure of the relative *impact performance*.

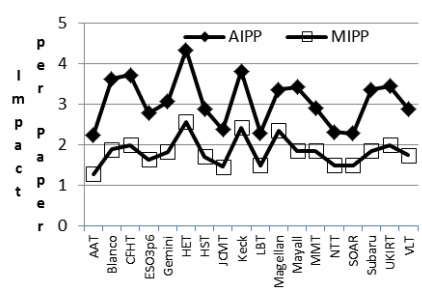


Figure 2. AIPP and MIPP of each telescope for the period 2007-2011

The AIPP and MIPP are shown in Figure 2 and the AIPP is significantly higher than the MIPP due to long tail of very high impact papers. The AIPP differs by a factor of approximately two between the lower performing telescopes and the higher performing ones.

A truly high performing telescope will combine high productively and a high average impact per paper, which is equivalent to a high *total impact*. The total impact of telescope's papers is simply the sum of the impacts of all the individual impacts of the papers published using data from that telescope.

The total impact of each telescope is displayed in Figure 3. The best performing telescope, Keck, combines a very high productivity with a very high AIPP. While the HET has the highest AIPP its low productivity means that it is one of the lowest performers.

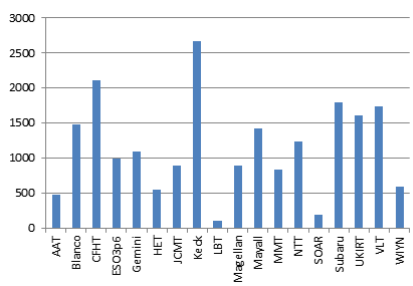


Figure 3. Total impact of each telescope for the period 2007-2011

Conclusions

A standard bibliometric approach provides a good measure of the comparative performance of modern telescopes. Normalizing the raw citation counts of papers to a standard measuring rod provides an age independent metric that can be used to compare publications of different ages. This approach should be used in any study that utilizes bibliometrics to study publications over a range of years.

Acknowledgments

This research has made use of NASA's Astrophysics Data System..

References

Crabtree, D.R. & Bryson, E.P. (2001). The Effectiveness of the Canada-France-Hawaii Telescope. *The Journal of the Royal Astronomical Society of Canada, 95, 259-265.*

Kurtz, M.J., Eichhorn, G., Accomazzi, A., Grant, C.S., Murray, S.S., Watson, J.M. (2000). The NASA Astrophysics Data System Overview. *Astronomy and Astrophysics*, 143, 41-59.

PROFILES OF PRODUCTION, IMPACT, VISIBILITY AND COLLABORATION OF THE SPANISH UNIVERSITY SYSTEM IN SOCIAL SCIENCES AND HUMANITIES

Daniela De Filippo[1]; Carlos García-Zorita[2]; Sergio Marugan[3] and Elías Sanz-Casado[4]

[1] *dfilippo@bib.uc3m.es;* [2] *czorita@bib.uc3m.es;* [3]*smarugan@pa.uc3m.es;* [4]*elias@bib.uc3m.es*

Department of Library and Information Sciences, Carlos III University of Madrid, 126 Madrid Street, Getafe 28903, Madrid(Spain)

Introduction

Despite the widespread use of international databases, the scarcity of publications from non-English speaking countries, as well as the lack of journals in Social Sciences and Humanities, has been highly criticized (Gómez and Bordons: 1996). This situation has created an absence of information on a production sector that is gradually gaining more and more visibility.

Between 2002 and 2011, Spanish publications in WoS increased from 30,088 to 61,364 documents, representing a 203% increase. In the same period, Spanish publications in Social Sciences and Humanities increased from a 7% of the country's total production to 15%, constituting an overall growth of 397,9%.

These data show the increasing significance of publications within these disciplines in WoS. This reality may be due either to the implementation of a specific dissemination strategy by researchers in order to increase their visibility, or to the fact that the requirements from the assessment agencies focus on the relevance of papers as key elements within the evaluation process..

Further evidence of the increasing visibility of these disciplines in WoS is shown by the number of journals. JCR data show that, in 2002, Spain had 26 SCI journals and only 2 SSCI journals, while in 2011 there were 78 SCI journals and 55 SSCI journals.

This increase can also be seen as a result of both the application of different projects such as REHS; MIAR and IN-RECs, which aim at analyzing and improving the quality of Spanish journals, and the efforts carry out by the publishers of Spanish journals to meet the requirements of international databases (Gimenez-Toledo: 2011).

To analyze in detail the evolution of Spanish publications in the Social Sciences and Humanities, we will study the production of the university system, which represents 67% of Spain's overall production. The main aims are:

- To identify activity profiles in these areas in terms of production, productivity, visibility, impact and collaboration.
- To analyze the annual evolution of each indicator.
- To compare the relationship between the intensity of activity in these areas and the specialization for each university.

- To analyze activity profiles in these areas vs. those defining the Spanish University System's total production.

Sources and Methodology

Data from the IUNE Observatory were used as source (Sanz-Casado et al: 2011). This Observatory (created for our research team) presents 42 indicators grouped in 6 dimensions with the aim of analyzing the activity of the Spanish University System (www.iune.es). In its current version, it provides disaggregated data by scientific areas.

The assignment of areas was carried out taking into account the WoS disciplinary classification of the journals, which have been grouped into six broad areas: Arts and Humanities; Life Sciences; Experimental Sciences; Architecture, Engineering and Computing; Medicine and Pharmacy and Social Sciences.

Publications from 49 public universities and 25 private institutions were identified through the use of a standardized normalization system, which has been developed by the Laboratory of Metrics Studies of Information (LEMI).

Scientific production from the Spanish University System for the period 2002-2011 was obtained. The following indicators for Social Sciences and Humanities versus the total System were calculated:

- Annual evolution of the number of publications.
- Publications per university.
- Annual evolution of the co-authorship rate.
- Annual evolution of the collaboration profiles.
- Citations received per university.
- Percentage of non-cited documents.
- Percentage of documents in Q1.
- Percentage of documents in TOP3 journals.

Results

Between 2002 and 2011, the Spanish University System has shown an important increase, from just over 20,000 publications in Web of Science to 41316. The area with the highest production was Experimental Science with 40% of the total (Fig.1).

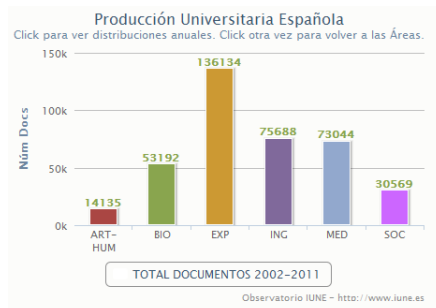


Figure 1. Distribution of Spanish Universities' Publications by Thematic Area

The data show that in the period being analyzed, the Spanish university system has had a total increase of 203% in its number of publications (similar to that of the entire country. Despite the fact that the area of Experimental Sci. has a higher production in absolute values, the greatest increase has taken place in the areas of Social Sci. and Humanities, with 274% and 227% respectively (Fig 2).

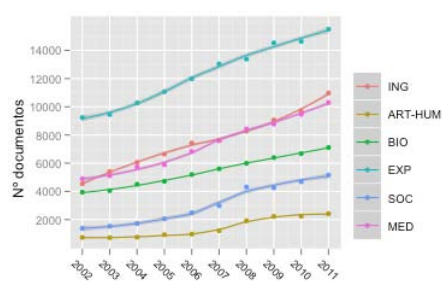


Figure 2. Annual Evolution of the Number of Publications in the Spanish University System by Thematic Areas

The production within the area of Social Sciences represents, on average, 11.4% of all Spanish universities' publications, while documents within the field of Humanities reach 5.2%. Figure 3 shows the distribution of production in each area for the 10 universities with the largest volume of documents. In these institutions, the percentage of publications in the Social Sciences and Humanities is lower than the system's average, as other smaller universities are the most important in these fields. At institutional level, the most relevant universities in the field of Social Sciences are: the National Distance Education University; Pablo Olavide University; Carlos III University of Madrid and Pompeu Fabra University with more than 20% of their total production. The most productive in Humanities are the National Distance Education University, the Complutense University of Madrid, Salamanca University and Alcala de Henares University.

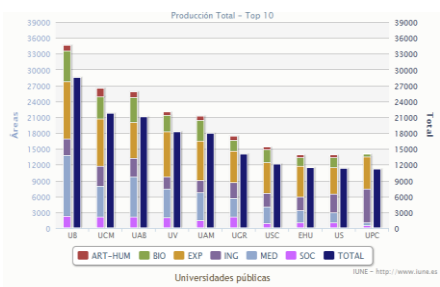


Figure 3. Distribution of Publications by Thematic Area in the Top10 Universities.

In the Spanish University System, the co-authorship rate was on average 5.3 author/doc in 2002, while it went up to 15.54 author/doc in 2011. However, these values are lower for the production levels in the Social Sciences and Humanities. The situation is the same in the case of documents in the first Quartil and for documents in the TOP3 journals.

On the contrary, the percentage of non cited documents is higher than the area average, particularly in Humanities. Taking into account indicators such as the collaboration profile, the Social Sciences show similar values to all areas, although Humanities have a different behavior than the rest of disciplines (Table 1).

Table 1. Activity Profiles in Social Sciences and Humanities versus All Areas

Indicator	2002	2011	Increase
All areas			
% Domestic Collaboration	29.5	34.30	16.27
% International Collaboration	34.2	39.80	16.37
Humanities			
% Domestic Collaboration	19.34	22.55	16.60
% International Collaboration	16.38	24.09	47.07
Social Sciences			
% Domestic Collaboration	30.72	32.02	4.23
% International Collaboration	28.79	32.14	11.64

Discussion

Throughout this paper, we show in detail each of the indicators analyzed. The activity profiles for each institution will be defined accordingly.

References

Giménez Toledo, E (2011). The Opinion of the Experts About Spanish Communication Journals and Other Quality Indicators. First National Convention *on Communication Research Methodology* (AE-IC and Rey Juan Carlos University). Fuenlabrada (Madrid), april 13th and 14th.

Gómez, I. and Bordons, M. (1996). Limitations in the Use of Bibliometric Indicators for Scientific Evaluation. *Política Científica*, 46: 21-16

Sanz Casado, E; De Filippo, D; García Zorita, C; Efraín-García, P. (2011). The IUNE Observatory: A New Tool for the Evaluation of Research Activity within the Spanish University System. *Revista BORDON,* 63 (2): 101-115

PROTOTYPICAL STRATEGY FOR HIGH-LEVEL CITATION-ANALYSES: A CASE STUDY ON THE RECEPTION OF ENGLISH-LANGUAGE JOURNAL ARTICLES FROM PSYCHOLOGY IN THE GERMAN-SPEAKING COUNTRIES

Günter Krampen[1], Gabriel Schui[2] and Hans P.W. Bauer[3]

[1] *krampen@uni-trier.de*
Leibniz Institute for Psychology Information and Documentation (ZPID) and University of Trier, Department of Psychology, D 54286 Trier (Germany)

[2] *schui@zpid.de*
Leibniz Institute for Psychology Information and Documentation (ZPID), D 54286 Trier (Germany)

[3] *bauer@zpid.de*
Leibniz Institute for Psychology Information and Documentation (ZPID), D 54286 Trier (Germany)

1 Introduction

The Web of Sciences (WoS) is frequently and increasingly used in evaluations of scientific productivity and of the international reception of publications (Garfield, 1979). However, citation-analyses done with this database are sensitive to problems and some dangers of misinterpretations, which result from the features of this database. Some of these problems refer to the limited representation of journal publications (because not all journals are documented in the WoS) and to name homonymy, i.e., identical names of different authors. Therefore, firstly, high-level citation-analyses should test the completeness of the papers documented in the WoS, secondly, citation-analyses must be implemented publication-based and not name-based.

A prototypical strategy for high-level citation-analyses considering and omitting both problems is presented in the following. The analyses refer to a case study on the international reception of English-language journal publications from psychology in the German-speaking countries (Austria, Germany, parts of Switzerland), which were published in three decades, 1981-2010. This prototypical strategy of citation-analysis includes five steps of data gathering, data management (including correction of data defects), and data analyses.

2 Five Steps of Data Gathering and Data Management

First Step: Identification of all Relevant Publications

For the identification of all relevant English-language journal publications from psychology in the German-speaking countries the exhaustive database for psychology publications from the German-speaking countries

PSYNDEX is used (http://psyndexdirect.zpid.de).
The source database PSYNDEXE includes in total 28,845 documentations of English-language psychology journal articles from the German-speaking countries for the publication years 1977-2011. PSYNDEX and STAR Record identification number (ID), author(s), publication year, title of the paper and of the journal, ISSN, volume, pages/number of paper, and DOI are registered for each of these articles. Eight faulty documentations in the database were eliminated (six doublets and two not identifiable documents), which led to a total of 28,837 articles.

Second Step: Assignment of Publications to WoS-UTs

To assure high precision the bijective assignments of PSYNDEX-IDs to WoS-UTs were carried out in three ways using the APIs of the WoS. Firstly, a query (API of the WoS Web Services, L1.02) including author names, publication year, and the three longest words of the title, resulted in the identification of 24,374 WoS-UTs of the 28,837 PSYNDEX-IDs (84,5%). Secondly, a query (WoS Links Article Match Retrieval, 1.4) including ISSN, volume and first page number or article number or author names resulted in the identification of 22,459 WoS-UTs (77,9%). Thirdly, the search for DOIs (documented in PSYNDEX) in the WoS resulted in 9,396 bijective assignments to WoS-UTs (32,6%). Different WoS-UTs were registered for 41 articles at least in two of the three assignment strategies. These cases were corrected manually without any problem. The small number of such assignment problems (0,2%) confirms the high reliability of the assignments by cross-validation.

To sum up, bijective (one-to-one) assignments to WoS-UTs were possible for 25,747 papers documented in PSYNDEX. This results in a WoS coverage of English-language journal publications from psychology in the German-speaking countries of 89,3%. Thus, approximately 10% of the articles documented in PSYNDEX are unconsidered in the WoS.

Third Step: General Citation-Analysis

On the basis of the WoS-UTs distinct citation frequencies of the 25,747 articles were gathered (July, 2012). Citation-Analyses referred to WoS segments SCI-E and SSCI, which allow—in the next step—analyses of citations per year (Garfield, 1979).

Fourth Step: Analyses of Citations per Year

Frequencies of citations per year were determined for all publications using WoS Web Services API. In addition to the registration of all citations received, the numbers of self-citations of authors were determined. This is significant because the number of self-citations varies strongly, i.e., it ranges from 0% up to 100%.

Fifth Step: Analyses of Features of the Articles Receiving Citations by Others

To get information about some characteristics of highly versus moderately versus rarely or not at all cited journal articles formal features of all 25,747 articles included in citation-analyses were determined with reference to PSYNDEX. These features include the sub-discipline of psychology, descriptors, and study type (e.g., methodological study, empirical/experimental study, theoretical study, literature review, overview etc.).

3 Results

Results refer to the English-language journal articles from psychology in the German-speaking countries, which were published between 1981 and 2010. PSYNDEX includes 28,388 article documentations for these three decades.

Bijective (one-to-one) assignments to WoS-UTs were possible for 25,461 of the papers documented in PSYNDEX (89,7%). This points at the fact that WoS coverage of English-language journal publications from psychology in the German-speaking countries is relatively satisfying, but not exhaustive. Approximately 10% of the articles are not considered in the WoS and therefore are excluded completely from citation-analyses. This can cause distinct biases in evaluations of the reception of the publications of authors and institutions.

The distributions of the absolute frequencies of citations of articles and of the frequencies of citations per year are strongly skewed to the right and resemble Pareto probability functions (Bauer et al., 2013; Seglen, 1992). Five-year impact factors (IF) increase continuously from 1980ties (IF = 1) to 2010 (IF = 3,5). This increase is observed for the reception of journal articles in all psychological sub-disciplines. However, the slope is different between the sub-disciplines being most steeply for neuropsychology and biopsychology.

Self-citations range between 0% and 100%. On average there are 17% self-citations with a slight decrease from the 1980ties (20%) to 2010 (16%).

Highest citation frequencies are registered on average for journal publications in clinical psychology and cognitive (experimental) psychology. It must be considered that both sub-disciplines are comparably large research community giving authors a greater chance of being cited.

Highest citation frequencies are registered on average for literature reviews and overviews, less for empirical and experimental studies, and fewest for methodological studies.

4 Conclusions

WoS coverage of English-language journal publication from psychological research in the German-speaking countries is relatively satisfying, but certainly not exhaustive. Approximately 10% of the publications are not included in the WoS, which can cause a serious selection bias in evaluative applications of citation-analyses.

Citation frequencies and five-year impact factors increase continuously from the 1980ties to 2010 showing distinct differences between the sub-disciplines of psychology. This is connected to the size of the research community in the sub-discipline.

Self-citations should be generally omitted in citation-analyses because of the large differences between authors and between publication years.

Literature reviews and overviews receive—on average—the highest citation numbers, empirical studies less, and methodological studies the lowest.

References

Bauer, H. P. W. Schui, G., von Eye, A., & Krampen, G. (2013). How does scientific success relate to individual and organizational characteristics? *Scientometrics,* 94, 523-539.

Garfield, E. (1979). Citation indexing: Its theory and application in science, technology, and humanities. New York: Sage.

Seglen, P. O. (1992). The skewness of science. *Journal of the American Society for Information Science,* 43, 628-638.

A QUANTITATIVE ANALYSIS OF ANTARCTIC RELATED ARTICLES IN HUMANITIES AND SOCIAL SCIENCES APPEARING IN THE WORLD CORE JOURNALS

Weina Hua[1] and Yu Li[2]

[1]*huawn@nju.edu.cn*
School of Information Management, Nanjing University, Nanjing 210093 (P. R. China)

[2]*baihe2010801@163.com*
School of Information Management, Nanjing University, Nanjing 210093 (P. R. China)

Introduction

The end of the earth south is the South Pole, Antarctica. As an under-developed, uncontaminated clean continent, it is the only natural laboratory for scientific research. Now, there are 100 and more Antarctic Stations established by more than 40 countries carrying out multi-disciplinary investigation, like life sciences, earth sciences, marine science, physic. Thus, it is of importance to have a good idea of the academic situation of Antarctic related research. In this contribution we focus on research in the humanities and social sciences, such as political sciences, geography and economics.

Methodology

Antarctic research focusing on the humanities and social sciences covers many disciplines. In order to have a better understanding and to evaluate the types of Antarctic research historically and to see current trends, a detailed analysis of Antarctic related articles appearing in the world core journals published between 1900-2011 was made. The data originate from the SSCI (Social Sciences Citation Index) and the A&HCI (Art & Humanities Citation Index), subfiles of the Web of Science (referred to WoS), collected from different aspects. Data was gathered by searching for Antarctic related subject headings in the title field including variant names such as Antarctic or South Pole or subantarctic. Antarctic Polar Regions such as King George Island, Alexander Island, Victoria Land, South George Island, Wedell Sea, Ross Sea. The names of Antarctic Stations such as McMurdo, Mawson, Halley were also included and organizations titled with Antarctica or South Pole were also in searching process. Search operators “and” ”or” “not” were used in the main search strategy. The study covers the time period from 1898-2012[185]. Data collection work ended on 1st of March 2013. After manual adjustment we got a total of 2121 hits.

Yearly Published Papers

The earliest papers in the world core journals relating to Antarctic retrieved from SSCI and A&HCI dated back to 1900. 2121 papers have been found out

[185] The database chosen for this research covers the data retrospectively from early 1898, but the data meeting the topics of this paper begins from 1900.

from 1900 to 2012 on Antarctic research in humanities and social sciences. The data in details are shown in Table 1.

Table 1. Number of articles each year

Year	*No. of articles*	*Year*	*No. of articles*
2012	45	1983	51
2011	37	1982	20
2010	39	1981	17
2009	27	1980	17
2008	45	1979	28
2007	37	1978	23
2006	31	1977	26
2005	25	1976	23
2004	27	1975	12
2003	23	1974	10
2002	27	1973	10
2001	37	1972	7
2000	44	1971	14
1999	28	1970	27
1998	25	1969	9
1997	16	1968	15
1996	28	1967	17
1995	27	1966	21
1994	46	1965	13
1993	34	1964	18
1992	26	1963	22
1991	40	1962	19
1990	48	1961	15
1989	48	1960	26
1988	38	1959	22
1987	34	1958	25
1986	47	1957	21
1985	37	1956	14
1984	49	1955	4
1954	10	1926	7
1953	5	1925	13
1952	10	1924	4
1951	4	1923	6
1950	10	1922	3
1949	2	1921	5
1948	8	1920	4
1947	8	1919	2
1946	1	1918	2
1945	2	1917	5
1944	4	1916	8
1943	3	1915	16
1942	1	1914	17
1941	4	1913	31
1940	12	1912	19
1939	15	1911	20
1938	8	1910	31
1937	10	1909	23
1936	9	1908	3
1935	9	1907	8
1934	7	1906	11
1933	12	1905	23
1932	12	1904	18
1931	9	1903	20
1930	16	1902	20
1929	12	1901	21
1928	5	1900	6
1927	8		

Fig 1 shows that in the first 20 years from 1900 to 1919 the research on Antarctic was relatively much more active than several decades years followed. Most of the works were on the category of geography. Document type as book review got much more shares compared with other years in this data collection. South Pole Expedition or exploration was the main theme of those years' research, such as British Antarctic Expedition 1898-1900; British Antarctic Expedition 1907-1909; British Antarctic Expedition 1910-13; German South Polar Expedition 1901-1903; German Antarctic Expedition 1911-1912; French Antarctic Expedition 1903-1905; French expeditions to the Antarctic 1908-10; Australasian Antarctic Expedition 1911-1914; Norwegian Antarctic Expedition 1910-1912; Swedish South Pole Expedition 1901-1903; Scottish National Antarctic Expedition during the Years 1902, 1903, and 1904; Belgian Antarctica Expedition; Bellingshausen's Expedition in Antarctica 1819-1821; Shackleton Antarctic Expedition; Amundsen's Antarctic Explorations. And some stations were also discussed at those early years, such as Argentine Antarctic Station and Kerguelen-Station.

No greater leap has been found during the following over 60 years from 1920 to 1982 when the yearly numbers of

articles on Antarctica was relatively small (less than 20 in most of the years). Of course polar expedition was still the topic during these years, but not the main theme then. The research scale was enlarged from geography to especially history, politics, economics, anthology, and so on. Naming, international cooperation, heroes, solitary were also taken into consideration. Antarctic Treaty was one of the topics emerged more often. The methodology was also enriched. Britain presented many examples of survey and investigation on Antarctic social issues. Not only the theoretical review but also the applied engineering or practical research was taken on, such as whaling, sailing, mapping, sight seeing, polar travel and so on, focusing on social factors since the data for this paper was limited in social sciences field. Review works, bibliographies, libraries, even Antarctic online databases have been mentioned during this period.

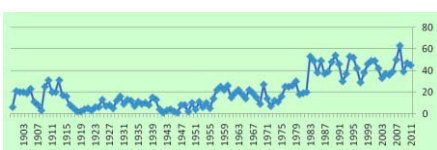


Fig 1. Growth curves for Antarctic studies in the humanities and social sciences

From 1983 to 2011 it took about one fourth of the time period, but produced more than half of the research papers showing more enthusiasm on Antarctic social science study by scientific scholars. Resource policy, Antarctic warming, environment protection, security issues, health problems were discussed. Research topics on ecology, psychology, psychiatry, law, became more attracting. Database systems of Antarctic information, budget for Antarctic has been mentioned several times. Tourism is also a topic coming in great numbers during this period.

Acknowledgments

Research for this article is funded by the Chinese Arctic and Antarctic Administration (Grants CHINARE2012-04-05-03-04).

References

Web of Science, available at: http://www.isiproducts.com

Qu, Tanzhou(2010). The history, Present situation, and future development. available at: http://www.szlib.gov.cn/newsshow.jsp?itemid=3114

Guo, Peiqing(2013). Polar regions in the future influences China a lot. available at: http://express.cetin.net.cn/cetin2/servlet/cetin/action/HtmlDocumentAction?baseid=1&docno=325689

THE RELATIONSHIP BETWEEN A TOPIC'S INTERDISCIPLINARITY AND ITS INNOVATIVENESS

Carolin Michels

carolin.michels@isi.fraunhofer.de
Fraunhofer Institute for Systems and Innovation Research, Breslauer Straße 48, 76139 Karlsruhe (Germany)

Introduction

There have been various attempts in the past to identify innovative or high impact topics in science semi or fully automatically. Many studies do this in retrospective and with the help of citations (see e.g. de Solla Price 1965) – a method that demands a time span of at least 2 years to give the scientific community enough time to discover, react and cite the topic in question. In that case, the identification of an innovative topic relies – on its basis - fully on the "wisdom of the crowd", i.e. the ability of the fellow researchers to discover and communicate the novel findings.

The goal of this study is to find out whether the interdisciplinarity of a topic might be used as an indicator to find topics that have a high innovation potential. Interdisciplinarity was defined as the combination of different fields or even topics in a field. The only other necessary condition in the case of topic combination is that the topics should developed so far independently and that therefore the combination of the topics is a novelty. We would argue that even though differences exist between multi-, inter- and transdisciplinarity (see e.g. Russel et al. 2008) the implications hold for all three kinds of combinations of knowledge across former boundaries.

The assumption that interdisciplinarity might be used as an indicator for innovation derives originally from Kuhn's definition of paradigm shifts (Kuhn 1970). The transfer or adoption of knowledge across boundaries can help to turn the corner in a crisis (Thompson-Klein 2004). For example, genetic algorithms use the basic biological principles of recreation and evolutionary survival of the fittest to facilitate complex mathematical calculations.

The combination of knowledge in turn can result in independent topics or fields (see e.g. Shafique 2013, Alvargonzalez 2011) that can evolve in an independent way. Sometimes, this might lead again to a diminishing multidisciplinarity, which might "hinder tapping the full potential of research" in severe cases (Shafique 2013).

Some findings already suggest that an interdisciplinary approach has more impact than monodisciplinary work. The impact (measured in whatever form) is a reasonable indicator for innovativeness. For instance, it has been shown that multi- or interdisciplinary work enhances the citedness (Leimu and Koricheva 2005, Levitt 2008) or the success rate (Sigelman 2009) and thus the impact of a paper. Albright argues that according to Adams (2006), creativity is a product of "the convergence of knowledge, creative

thinking, and motivation" (Albright 2010) and since these factors are promoted by multidisciplinary work, multidisciplinarity leads to creativity which in turn causes innovation.

Data and Methodology

The document set was extracted from an in-house implementation of Elsevier's Scopus database. All articles in the category AI (category 1702) that had a title and an abstract, at least 5 references and at least 2 citations were collected and 1,000 documents per year were selected randomly. Because of the lower data coverage in the years 2010 and later, we restricted our data analysis to the years 2000 to 2009. Only non-Computer Science citations were classified as interdisciplinary citations.

Latent Dirichlet Allocation (LDA) was used to generate the topic clusters in the beginning (Blei et. al 2003). We extended the LDA model to take into account references, as has been done in similar studies before (Erosheva et al. 2004, Nallapati et al. 2008).

Interdisciplinarity was measured via citations and references for the documents. For the disciplines, we analyzed those clusters for which more than 50% of the citations were emitted by documents in other disciplines or that cited at least in 50% of the cases documents in other disciplines. Concerning the topics, we analysed those clusters that cited two clusters which had never been cited together before.

Results

Disciplines

There were many clusters that had a high citation rate of other disciplines because the respective documents actually belonged to these other disciplines. This was particularly the case for high citation rates of Chemistry. However, there were also some topics that cited other disciplines extensively because they adopt or transfer knowledge, e.g. aspects that are transferred from human behaviour to AI. In addition, some clusters refer to Biochemistry or Medicine because AI is applied to model biological processes.

Thus, the interdisciplinarity of the reference list (rather than the citation list) might indicate a high level of innovativeness. However, this indicator might be misleading as in the examples shown above.

Table 1. Triples of cluster, where the citing cluster is the first to cite the two clusters together.

	Cluster (Year)	Cited Cluster 1	Cited Cluster 2
1a	"robust" output control/ performance, nonlinear systems (2009)	Stability analysis	Teleoperations and autonomous vehicles
1b		Fuzzy/feed-forward control	Identification of dynamic systems
2	Sparse (Bayesian) modeling, dimensional reduction (2009)	Sampling/ experiment selection for face/object/ image recognition	Independent component analysis
3	Classifier ensembles (2007)	Face recognition/ classification	Inconsistencies in structured text
4	Robot navigation, path planning (2008)	Target tracking with robots	Fuzzy behaviour robots and multiple object tracking
5	Genetic algorithms (2008)	Dimensional knowledge reduction	Image transformation

Contexts

Table 1 shows the 6 triples of citing and cited clusters where the citing cluster connected so far unconnected clusters.

We cannot discuss the details of the clusters here, but just summarize that in terms of innovativeness, all but triple 4 seem to be highly innovative. Even if the aspects of a cluster are not new themselves, the approach of combining so far unrelated topics for the old task might be innovative (as for example done in triple 3 or 5).

Conclusion

In previous work, the relationship between field dynamics, innovation and interdisciplinarity has been studied in one direction, i.e. it was shown that high innovative topics had a high interdisciplinarity. In this work, we tried to find topics with a high innovativeness via their interdisciplinarity and use of other topics. The results suggested that the topics that were cited by different other topics were highly volatile, ambiguous or dynamic but not necessarily innovative. Being cited by other fields did not necessarily indicate a high innovativeness. However, the manual assessment of the clusters citing new combinations of topics showed that indeed these clusters were in most cases high impact clusters.

References

Adams, K. (2006): The sources of innovation and creativity. Washington, DC: National Center on Education and the Economy.

Albright, Kendra (2010): Multidisciplinarity in Information Behavior: Expanding Boundaries or Fragmentation of the Field? In: *Libri* 60 (2).

Alvargonzález, David (2011): Multidisciplinarity, Interdisciplinarity, Transdisciplinarity, and the Sciences. In: *International Studies in the Philosophy of Science* 25 (4), pp. 387–403.

Blei, David M.; Ng, Andrew Y.; Jordan, Michael J. (2003): Latent Dirichlet Allocation. In: *Journal of Machine Learning Research* (3), pp. 993–1022.

de Solla Price, Derek J. (1965): Networks of Scientific Papers:The pattern of bibliographic references indicates the nature of the scientific research front. In: *Science Communication* 149 (3683), pp. 510–515.

Erosheva, Elena; Fienberg, Stephen; Lafferty, John (2004): Mixed Membership Models of Scientific Publications. In: *Proceedings of the National Academy of Sciences of the United States of America, 101 (Suppl 1),* pp. 5229 - 5227.

Kuhn, Thomas S. (1970): The Structure of Scientific Revolutions. Second Edition, Enlarged: The University of Chicago Press.

Leimu, Roosa; Koricheva, Julia (2005): Does Scientific Collaboration Increase the Impact of Ecological Articles? In: *BioScience* 55 (5), pp. 438–443.

Levitt, Jonathan M.; Thelwall, Mike (2008): Is multidisciplinary research more highly cited? A macrolevel study. In: *J. Am. Soc. Inf. Sci.* 59 (12), pp. 1973–1984.

Nallapati, Ramesh; Ahmed, Amr; Xing, Eric P.; Cohen, William W. (2008): Joint Latent Topic Models for Text and Citations. In: *KDD'08*, pp. 542–550.

Russell, A. Wendy; Wickson, Fern; Carew, Anna L. (2008): Transdisciplinarity: Context, contradictions and capacity. In: *Futures* 40 (5), pp. 460–472.

Shafique, Muhammad (2013): Thinking inside the box? Intellectual structure of the knowledge base of innovation research (1988-2008). In: *Strat. Mgmt. J.* 34 (1), pp. 62–93.

Sigelman, Lee (2009): Are Two (or Three or Four ... or Nine) Heads Better than One? Collaboration, Multidisciplinarity, and Publishability. In: *PS: Political Science & Politics* 42 (03), pp. 507.

Thompson-Klein, Julie (2004): Prospects for transdisciplinarity. In: *Futures* 36 (4), pp. 515–526.

HIERARCHICAL CLUSTERING PHRASED IN GRAPH THEORY: MINIMUM SPANNING TREES, REGIONS OF INFLUENCE, AND DIRECTED TREES.

Xavier Polanco

xavier.polanco@gmail.com
IU Independent Unit, 11 rue Meslay, 75003 Paris (France)

Introduction

The clustering presented in this document includes algorithms based on graph theory, such as the Minimum Spanning Tree, and its extension that involves Regions of Influence, and also Directed Trees.

The question is how to do clusters with graphs. In the first section, the simplest case is considered; the second section deals with a more elaborated issue demanding a combination of methods; the last section refers to the example of directed graphs.

Minimum Spanning Trees

A Spanning Tree (ST) is a connected subgraph that is a tree containing all the vertices of the graph and having no loops, i.e., there exists only one path connecting two pairs of nodes in the graph. If the edges of the graph are weighted, the weight of the spanning tree is defined as the sum of the weights of its edges. A single graph can have many different spanning trees. In addition, a spanning tree as any graph can be labelled making the assumption that the semantic meanings of interest can be represented as such.

A Minimum Spanning Tree (MST) is the spanning tree with the smallest weight among all spanning trees connecting the nodes of the graph. There may be more than one minimum spanning tree for a given graph. Algorithms to define the MST of a graph: Florek et al (1951), Kruskal (1956) and Prim (1957).

Suppose a data matrix X of n cases or observations and p variables or attributes or patterns, and we desire obtain a clustering. For this, a proximity matrix is defined, P(X), which is $n \times n$, that is, a squared and symmetric matrix. From P(X) is induced a graph G(X), with $V = n$ vertices and a set E of edges with a weight w which is defined by the metric of P(X). The next step is determining the MST of the G(X). Finally, the clusters are the connected components of the MST, after the removal of the edges with the largest lengths compared with their neighboring edges.

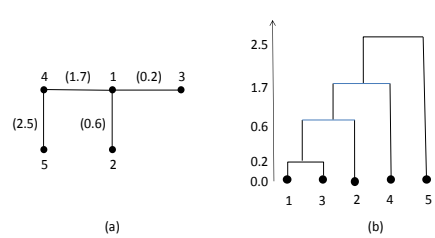


Figure 1. (a) The MST derived from a dissimilarity matrix. (b) The dissimilarity dendrogram obtained with the MST algorithm.

Once the MST has been determined using any suitable algorithm, we may identify a hierarchy of clusters, which is identical to the one defined by the single link algorithm, at least for the case in which all distances between any two vectors of X are different from each other. Thus, this MST may be viewed as an alternative implementation of the single link algorithm.
Observe that a MST uniquely specifies the dendrogram of the single link algorithm.
For the use of the MST for the cases of touching clusters, or clusters with different densities, see Zahn (1971).

Regions of Influence

An extension of the MST involves Regions of Influence (ROIs), noted $R(x_i,x_j)$. The algorithms for clustering applications are known as Gabriel Graph (GG) (Gabriel & Sokal, 1969; see also Matula & Sokal, 1980) and Relative Neighborhood Graph (RNG) (Toussaint, 1980; see also Urquhart, 1982, and Jaromczyk, & Toussaint, 1992). The idea of ROIs has been used in order to overcome the problems associated with the MST algorithms.

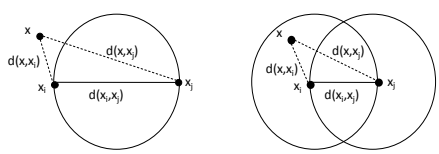


Figure 2. (a) Diagram of the Gabriel Graph. (b) Diagram of the Relative Neighborhood Graph.

A Region of Influence is defined as:

$$R(x_i, x_j) = \left\{x: cond\left(d(x,x_i), d(x,x_j), d(x_i x_j)\right), x_i \neq x_j\right\}$$

Where
$cond\ (d(x,x_i), d(x,x_j), d(x_i x_j))$ is a condition among the distances $d(x,x_i)$, $d(x,x_j)$, and $d(x_i x_j)$. Different choices of *cond* give rise to different shapes of regions of influence. In Gabriel & Sokal (1969) and Toussaint (1980), the following two choices are proposed:
$max\{d(x,x_i), d(x,x_j)\} < d(x_i x_j)$; and
$d^2(x,x_i) + d^2(x,x_j) < d^2(x_i x_j)$.

GG and RNG are similar in that the Minimal Spanning Tree is a subgraph of each. Thus, from the GG we can compute its RNG, and from the RNG we can compute its Minimum Spanning Tree (MST). Previously a Delaunay triangulation (DT) is applied in (Jaromczyk, & Toussaint, 1992). For any finite set X of points the following relations hold:
$MST(X) \subseteq RNG(X) \subseteq GG(X) \subseteq DT(X)$

The process to follow when the MST cannot be directly induced from a graph G(X) is: DT(X) → GG(X) → RNG(X) → MST(X). For an example, see Stout et al (2009), the authors apply this process in a domain of bioinformatics. In this way we can unravel a tangled graph as frequently occurs when X is a big data matrix.

Directed Trees

Many relations are directional, i.e. the ties are oriented from one actor to another. The citation composed by citing/cited is an example of a directional relation. In these cases, the graph is e directed graph or digraph. Thus, the cluster analysis consists to identify the directed trees of a digraph so that each directed tree corresponds to a cluster. A clustering algorithm is proposed in (Koontz et al, 1976). The resulting clusters are unimodal sets.
For each point x_i its neighborhood is defined:

$$\rho_i(\theta) = \{x_j \in X: d(x_i,x_j) \leq \theta, x_j \neq x_i\}$$

Where θ determines the size of the neighborhood and $d(x_i, x_j)$ is the distance between the corresponding points of X. Let $n_i = |\rho_i(\theta)|$ be the number of points of X lying in $\rho_i(\theta)$; and $n_i = (n_j - n_i)/d(x_i, x_j)$ will be used to determine the position of the point x_i in a directed tree.

The root x_i of a directed tree has the largest n_i among the points lying in $\rho_i(\theta)$; x_i is the point with the most dense neighborhood. It should be pointed out that this algorithm is sensitive to the order in which the vectors are processed.

Conclusion

The agglomerative single-link cluster analysis based on graph theory may then be phrased according to Minimal Spanning Tree model (Gower & Ross 1969; Dubes & Jain 1980; Hartigan 1985; Lebart et al, 1995) and its extension called Regions of Influence, i.e. Gabriel Graph and Relative Neighborhood Graph. When the graph is a directed graph the idea is the identification of directed trees in a digraph.

This document proposes to translate these techniques to scientometrics domain. The road was open long time ago in scientometrics by Okubo et al (1992) associating Correspondence Factorial Analysis and Minimum Spanning Tree (see also Miquel & Okubo 1994).

References

Dubes, R. & Jain, A. (1980). Clustering methodologies in exploratory data analysis. *Advances in Computers*, 19, 113-228.

Florek, K., Lukaszewicz, J., Perkal, J. & Steinhaus J. (1951). Sur la liaison et la division des points d'un ensemble fini. *Colloquium Mathematicum*, 2, 282-285.

Gabriel, K. R. & Sokal, R. R. (1969), A new statistical approach to geographic variation analysis, *Systematic Zoology* (Society of Systematic Biologists), 18 (3), 259–270.

Gower, J.C. & Ross, G.J.S. (1969). Minimum spanning trees and single linkage cluster analysis. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 18 (1), 54-64.

Hartigan, J.A. (1985). Statistical theory in clustering. *Journal of Classification*, 2, 63-76.

Jaromczyk, J.W. & Toussaint, G.T . (1992). Relative neighborhood graphs and their relatives. *Proceedings of the* IEEE, 80 (5), 1502-1517.

Koontz, W.L.G., Narendra, P.M. & Fukunaga, K. (1976). A graph theoretic approach to nonparametric cluster analysis. IEEE *Transactions on Computer*, 25 (9), 936-944.

Kruskal, J.B. (1956). On the shortest spanning subtree of a graph and the traveling salesman. *Proceedings of the American Mathematical Society*, 7 (1), 48-50.

Lebart, L., Morineau, A. & Piron, M. (1995). *Statistique exploratoire multidimensionnelle*. Paris, DUNOD.

Matula, D. W. & Sokal, R. R. (1980). Properties of Gabriel graphs relevant to geographic variation research and clustering of points in the plane. *Geographical Analysis*. 12 (3), 205–222.

Miquel, J.F. & Okubo, Y. (1994). Structure of international collaboration in science. 2. Comparison of profiles in countries using a link indicator. *Scientometrics*, 29 (2), 271-297.

Okubo, Y., Miquel, J.F., Frigoleto, O., & Doré, J.C. (1992). Structure of international collaboration in science: topology of countries

through multivariate techniques using a link indicator. *Scientometrics*, 25 (2), 321-351.
Prim, R.C. (1957). Shortest connection matrix network and some generalizatios. *Bell System Technical Journal.*, 36, 1389-1401.
Stout, M., Bacardit, J., Hirst, J.D., Smith, R.E., & Krasnogor, N. (2009) Prediction of topological contacts in proteins using learning classifier systems. *Soft Computing*, 13 (3), 245-258.
Toussaint, G.T. (1980). The relative neighborhood graph of a finite planar set. *Pattern Recognition*, 12 (4), 261-268.
Urquhart, R. (1982). Graph theoretical clustering based on limited neighborhood sets. *Pattern Recognition*, 15 (3), 173-187.
Zahn, C.T. (1971). Graph-theoretical methods for detecting and describing gestalt clusters. IEEE *Transactions on Computers*, 20 (1), 68-86.

RESEARCH SECTORS INVOLVED IN CUBAN SCIENTIFIC OUTPUT 2003-2007

Ricardo Arencibia-Jorge[1], Elena Corera-Alvarez[2], Zaida Chinchilla-Rodríguez[2] and Félix de Moya-Anegón[2]

[1] *ricardo.arencibia@cnic.edu.cu*
National Center for Scientific Research, 25 Avenue and 158 Street, Cubanacan, Playa, AP 6414, Havana (Cuba)

[2] *felix.demoya@scimago.es*
SCImago Research Group, Institute of Public Policies and Goods, CCHS-CSIC, Albasánz 26-28, Madrid (Spain)

Introduction

The analysis of scientific output and relationships of the different research sectors involved in Science and Technology policies using bibliometric methods is always a complex task.

Countries and regions have their own particular characteristics. The inclusion of any kind of institution in a specific sector requires the previous study of national science systems, and it depends on the objectives and functions of each institution in the national or regional environment. On the other hand, the behavior of inter-sector relationships is also strongly related with principles and norms of national science policies.

Research sectors involved in Cuban scientific activity are not yet fully studied. A previous paper of the authors explores the Cuban output at macro level (Arencibia-Jorge & Moya-Anegón, 2010). The current work analyzes the scientific activity and impact of the different Cuban research sectors during the period 2003-2007.

Method

Scopus was chosen as data source. A search strategy based on the identification of the word "Cuba" in *Author Address* and *Affiliation Country* fields was used. The retrieved items were downloaded to an *ad hoc* database, with the aim to eliminate false items and normalize affiliation data. Cuban research sectors were the most important aspect identified in each register from the database.

Data was collected in January 2010. The Scopus retrospective coverage process does not significantly affect the comparison between data obtained in the current work and those obtained from the earlier paper based on the same period.

The scientific production was distributed in six sectors: Higher Education (HEd), Health, Science & Technology (S&T), Government Administration (GAd), Enterprises (EPr) and Others.

Total publication output (A) and annual percentages were the indicators selected to show the quantitative dimension of the scientific production.

The qualitative dimension was studied through a set of impact indicators: Total of cited articles (AC), percentage of cited articles, average of citations per article (Ave) and H index (H). Each of these impact indicators were calculated by sector.

Social Network Analysis (SNA) techniques were employed to visualize the inter-sector collaboration.

Results and Discussion

The Cuban scientific output is mainly distributed in three of the six sectors analyzed (table 1). Higher Education is the most productive sector (55.4 %), which has been observed in previous studies (Sancho *et al*., 1993; Araujo-Ruíz *et al*., 2005).

In terms of visibility, Higher Education is the sector with the highest H-index, although their proportion of cited articles and the average of citations per article are below the national mean. The cause of this behavior is the big amount of papers published in less cited national journals, an aspect that involves the output of higher institutes of Medical Sciences belonging to the Higher Education sector, and hospitals belonging to the Health Sector.

Only 43 % of articles published by universities, and 31 % of articles published by hospitals were cited during the period, in contrast to the citation activity of Science and Technology (54 %) and Government Administration (55 %).

Government Administration comprised only 3.3 % of the national output, but showed the highest average of citations per article. Meanwhile, the sector Enterprises only published 71 articles during the period, with no role for any particular institution, and poor visibility.

Table 1. Scientific output and impact by sector.

Sector	A	%	CA	%	C	Ave	H
HEd	3199	55.4	1384	43.3	7680	2.40	29
Health	2270	39.3	708	31.2	4237	1.87	24
S&T	1864	32.3	1012	54.3	5521	2.96	25
GAd	190	3.3	105	55.3	777	4.09	17
EPr	71	1.2	19	26.8	62	0.87	4
Others	51	0.9	6	11.8	16	0.31	2
Cuba	*5778*	*100*	*2582*	*44.7*	*14727*	*2.55*	*34*

Institutions belonging to the scientific park from the west of Havana were the leaders of the sector Science and Technology. This group of institutions is responsible for an increasing amount of cash income that has made Cuba's biotechnology industry the third motor of the country's economy at the end of the decade.

The growth of Higher Education and Health Sector determines practically the nation's growth in global terms.

The scientific collaboration expressed in the different sectors analyzed can be approached from multiple perspectives. On the one hand, the collaboration among sectors offered an important view of the national scientific activity. On the other one, the establishment of strong networks of international collaboration was a very important strategy with the aim to achieving a high visibility or impact.

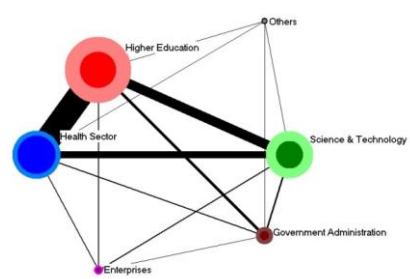


Figure 1. Inter-sector collaboration (UCINET 6.123; NetDraw 2.38).

On the figure 1, the size of the nodes identifies the volume of the sector´s output, the node ring represents the proportion of international collaboration, the lines imply the existence of collaboration among sectors, and the thickness of links expresses the intensity of those relations. Thus, the structural dimension of the national scientific output from the

characterization of its strategic sectors was objectively represented.

There are some problems that can be inferred from the presented map. First, there is a weak linkage between universities and institutions of science and technological innovation, and also reduced relationships between scientific research centers and health institutions in the country; second, the international collaboration is not representative in the health sector, taking into account the many Cuban health experiences and missions throughout the world; and third, there is a divorce between R&D units of enterprises and the institutions belonging to Higher Education and Science and Technology, given by the still insufficient research activity generated by Cuban enterprises.

The causes of these problems have a multifactorial nature. Despite the advanced research policy of higher education in Cuba, it is evident that still are low the level of actions developed by scientific institutions with the aim to attract the interest of students and research teams from universities. In this sense, the necessary link between the academy and institutions of science and technological innovation must be more evident.

On the other hand, taking into account the wide biomedical scope of Cuban scientific activity, there is no reason to avoid the collaboration between hospitals, health care centers and research institutions in research processes. An important number of Cuban products developed by scientific research centers are introduced in hospitals and distributed by the national network of pharmacies. Therefore, a more active role of physicians and professors from hospitals and health institutions, especially in research lines related to the use of these products by Cuban population, is necessary. Finally, it is clear that the absence of incentives is the main cause of a low international collaboration in the health sector, as well as the complete divorce between Cuban enterprises and institutions belonging to Higher Education and Science and Technology. In this sense, the recent creation of a biotech company (BioCubaFarma) that involve the most important research centers from the west of Havana, is a decisive step of the country in order to change the current status.

Conclusions

Cuban scientific output has experienced increasing growth during the first decade of the new millennium. The country´s efforts and expenditures in Research and Development activities had positive implications for the Cuban science system evolution, and total output of the country is led by the research developed in institutions belonging to the most relevant sectors involved in scientific activities: Higher Education, Health, and Science & Technology. Inter-sector relationships reveal some weaknesses in the national scientific macro-structure. Therefore, a deep analysis of the science and technology policies in each of the sectors studied is still necessary.

Acknowledgments

To the staff of *SCImago Research Group*, for the support, research advices and unconditional friendship.

References

Arencibia-Jorge, R. & Moya-Anegón, F. (2010). Challenges in the study of Cuban scientific output. *Scientometrics*, 83, 723-737.

Sancho, R., Bernal, G., & Gálvez, L. (1993). Approach to the Cuban Scientific Activity by Using Publication Based Quantitative

Indicators (1985-1989). *Scientometrics*. 28, 297-312.

Araújo-Ruiz, J.Á., Torricella-Morales, R.G., Van Hooydonk, G., & Arencibia-Jorge, R. (2005). Cuban scientific articles in ISI citation indexes and CubaCiencias databases (1988-2003). *Scientometrics*. 65, 161-171.

RESEARCH TRENDS IN GENETICS: SCIENTOMETRIC PROFILE OF SELECTED ASIAN COUNTRIES

Shivappa L. Sangam[1], Uma B. Arali[1], Chandrashekhar G. Patil[2] and Srishail Gani[3]

slsangam@yahoo.com

[1] Department of Library and Information Science, Karnatak University, Dharwad-580003. India.

[2] Department of Applied Genetics, Karnatak University, Dharwad-580 003. India

[3] Department of Statistics, Karnatak College, Dharwad-580 001. India

Introduction

Genetics a discipline of biology is the science of genes, heredity, and variation in living organisms. It is one of the youngest and the fastest growing disciplines of science. Now-a-days, it is a multidisciplinary subject as genes are universal to living organisms and genetics can be applied to the study of all living systems, from viruses and bacteria, through plants and domestic animals to humans. Knowledge of genetics being basic to progress in biology, agriculture, medicine, biotechnology, forensic sciences and many other fields, results of such studies are found highly useful. Research plays a vital role for the development or growth of subject(s) both qualitatively and quantitatively. This change in the trend can be traced by generating numerical data on the basis of the empirical evidence available in the process of identifying the trends in research priorities in a field. In the recent past, studies dealing with the assessment of scientific research in genetics by different nations have been reported in literature.

Methodology

For the present study, the data has been collected from PubMed database. It is one of the popular source of information available with NCBI (National Centre for Biotechnology Information) on the website addressed http://www.ncbi.nlm.nil.gov/pubmed/. The study compares the research priorities of 16 sub-specialities of genetics in 10 countries for two time-spans; 1992-2001 and 2002-2011.

1. Publication output of World in 16 sub-specialities of Genetics

From the table 1 and fig.1 it is evident that, Molecular Genetics and Human Genetics accounts for 68% of the total output in 1992-2001 and 60% in 2002-2011. Molecular Genetics accounts for the largest output 38% in 1992-2001 and Human Genetics 30% in 2002-2011 block periods. Ecological Genetics accounts for the smallest output of 0.1% in 1992-2001 and Genetics of Intelligence 0.2% in 2002-2011. The table clearly indicates the importance of trends in all the 16 branches which has increased or decreased over a period of twenty years.

2. Publication output and share of publications of major Asian countries

The publication output and share of publications of major nations of Asian continent is shown in table 2 and also is represented in form of graph in figure 2.

Among fifty Asian countries, Japan alone accounts for about 66% among Asian output in 1992-2001 and Japan and China 65% in 2002-2011.

Table-1 Publication output of world in 16 sub-specialities of genetics

	Number of Publications			
Branches	**1992-2001**		**2002-2011**	
	World	**%**	**World**	**%**
B G	5836	0.63	14224	0.84
Cl G	6128	0.66	12821	0.76
D G	30800	3.32	62457	3.7
C G	6458	0.7	12854	0.76
E G	1000	0.11	6775	0.4
Ev G	9861	1.06	31090	1.84
G E	54710	5.9	88720	5.26
G I	1420	0.15	3067	0.18
G	8928	0.96	59080	3.5
H G	277028	29.88	498481	29.56
M G	109206	11.78	198939	11.8
Mi G	12787	1.38	33730	2
Mo G	350537	37.81	506201	30.02
P G	35015	3.78	91761	5.44
Ps G	2790	0.3	6916	0.41
Q G	14632	1.58	59160	3.51
Total	927136	100	1686276	100

Note: B G-Behavioural Genetics, Cl G-Classical Genetics, D G- Developmental Genetics, C G-Conservation Genetics, E G-Ecological Genetics, Ev G-Evolutionary Genetics, G E-Genetic Engineering, G I-Genetics of Intelligence, H G-Human Genetics, M G-Medical Genetics, Mi G-Microbial Genetics, Mo G-Molecular Genetics, P G-Population Genetics, Ps G-Psychiatric Genetics, Q G- Quantitative Genetics.

3. Publication performance of Major countries

The data in table 4 reveals articles published in 16 branches, represented in different columns and 10 countries in rows during 2002-2011. In the present block period also, it is Japan which has maximum number of articles published (130060) however, it is followed by China (115689); India (26456); Israel (22033); Taiwan (21970); Georgia (9421); Turkey(7412); South Korea(6498); Hong Kong(5968); and Russia(5832).

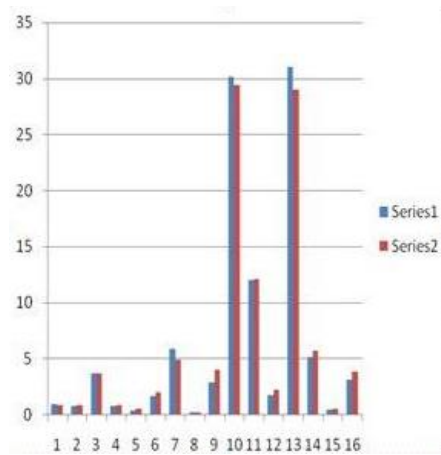


Fig. 1. Growth comparison between sub-specialities of Genetics

4. Priority profiles of Major countries:

The raw count alone does not convey much information as these figures are confounded by the size of the countries and the size of the subject specialities. Hence, an index called **Relative Priority Index (PI)** is computed for cross national comparison using formula as suggested by Nagpaul and Pant [6].

$$(PI)_{(i,j)} = \frac{nij/_{ni.}}{n.j/_{n..}} X\ 100$$

The profile of research priorities of major countries are presented in Table 5 and 6 for two time-spans- 1992-2001 and 2002-2011. These tables indicate the differences in the priority accorded to different sub-specialities by different countries.

From the values of PI, we can compare:

i) The research priorities of a country for different sub-specialities of genetics in a given time span;

ii) The research priorities of different countries for a given sub-speciality in a given time span; and
iii) The research priorities of a country for a given sub-speciality at different time span.

In table 5 and 6, the PI value for each variable in the table 3 and 4 is computed using PI formula i.e. for both the block period 1992-2001 and 2002-2011. The priority index of different countries is arranged again in the form of a matrix where the rows represent the countries and columns the sub-specialities. So, the row vector represents the priority profiles of countries, whereas the column vectors geographical profiles of sub-specialities.

Conclusion:

Genetics, the most rapidly growing area of research has relevance to many aspects of human life and society, including health, behaviour, food production, forensics and even politics. If used wisely, the new information promises to enhance our quality of life. Science indicators are used for both descriptive as well as analytical purposes to identify trends, make comparisons and as an aid for theoretical understanding of casual structure related to science and technology systems.

On the basis of the analysis the following conclusions may be drawn -.

1. Among the 10 major Asian countries, Japan accounts for the largest output of genetics literature. It alone accounts for 66% of the Asian output in 1992-2001 and 38% in 2002-2011.
2. Among fifty Asian countries, India occupies third position in contributing to the genetics research.
3. Among sub-specialities, Molecular genetics accounts for the largest output of 38% in 1992-2001and 30% in 2002-2011.
4. The sub-speciality, Genetics of Intelligence accounts for the lowest output of 0.2% during both block periods.
5. PCA analysis indicates three groups of sub-specialities based on correlation among them.

References

Garag,K.C. at el. Scientometric profile of genetics and heredity research in India.*Annals of Library and Information Studies*.57, 2010,196 - 206.

Garg, K. C. at el. Plant Genetics and breeding research: Scientometric profile of selected countries with special reference to India. *Annals of Library and Information Studies*. 58, 2011, 184-197.

Varaprasad, S. J. D. and Ramesh, D. B. Activity and growth of chemical research in India during 1987-2007. *DECIDOC Journal of Library and Information Technology,* 31 (5), 2011.

Sangam, S. L. at el. Indicators for Demographic Research: A Cross-national Assessment. *Journal of Library and Information Science*. 30, 2005, 1-19.

Sangam, S. L., Arali, U. B., Patil, C. G. and Mageri, M. N. Scientometric Analysis of Genetic Literature. Proceedings of 8th International Conference on Webometrics, Informatics and Scientometrics & 13th COLLNET Meeting, October 23-26, 2012, Seoul, South Korea.

Nagpaul, P.S. and Pant, Nagaraj. Research priorities of major countries in artificial intelligence. Proceedings of International Congress on Cybernetics and Systems. NewDelhi. Tata McGraw Hill, 1993.

THE RISE AND FALL OF GREECE'S RESEARCH PUBLICATION RECORD: THE LAST 30 YEARS

Olga Koukliati[1] Anastassios Pouris[2]

[1] *julieparos@yahoo.gr*
Institute for Technological Innovation, University of Pretoria, Pretoria (South Africa)

[2] *apouris@icon.co.za*
Institute for Technological Innovation, University of Pretoria, Pretoria (South Africa)

Introduction

Monitoring the performance of national scientific and technological systems became a preoccupation of policy authorities internationally in the 1990s. (Pouris A, 2003). This poster identifies Greece's research publication record across major disciplines to the year 2010 and traces the various policies that affected the country's research system. The indicators covered are number of publications, citations received and their distribution across scientific fields and institutions. Bibliometric indicators are used internationally to monitor the outputs of scientific systems, they are clearly defined and unambiguous and allow categorization in particular scientific fields and disciplines (Pouris A, 2012). Such categorization is useful for judging the performance of national scientific systems and their component parts in a global context (Pouris A, 2003). Scientometric indicators are used widely as an indispensable part of science and technology policy monitoring and assessment studies (Jeenah M & Pouris A, 2008) related to the structure and dynamics of science, impact assessments and others (King D.A, 2004). Such indicators allow comparisons of different disciplines between countries and other metrics which are not possible through other methods. (Pouris A, [2]2012). The following approach is based on data provided by The National Documentation Centre of Greece (NDC/EKT-2012) and summarizes the information available in the databases of the National Science Indicators (NSI) and NCR of Thomson Reuters, with regard to Greece's number and share of the world's publications.

Findings: Greece's Research Performance

Greek publications, regarding the appeal, originality, quality and name recognition were better positioned internationaly in recent years: the aggregate indices and Greece's position internationally upgraded, impact of publications increased and performance of organizations improved. Growth was continuous until 2008, while in 2009-2010 the number of Greek papers downturned. In 2009 the continuous upward trend stopped. Greece follows the average performance of OECD and EU countries and records an almost zero coefficient of variation. In 2010 the decline in the number of publications in OECD and EU countries includes Greece.

Figure 1 shows the number of Greece's publications for the period 1981-2011. According to the NSI database, there

were 10.219 Greek publications in international scientific journals registered in the Web of Science in 2010. Greece's yield of research publications shows decreasing trends and slipped from 10.625 publications in 2008 to 10.579 in 2009 and 10.219 in 2010.

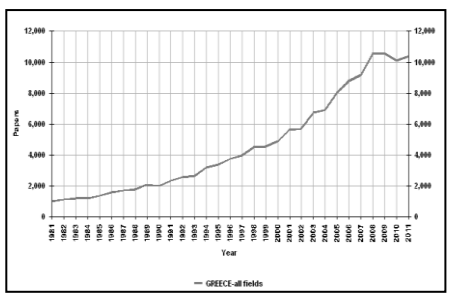


Figure 1: Number of Greece's overall publications

Based on 2010 figures, Greece participates in 2.4% of scientific publications in the EU and 1.14% in the OECD, doubling the units compared with 1996, and is in 24th among the 34 OECD countries.

In 2010 most Greek publications (48.9%) are classified in the scientific fields of *Physical Sciences, Medical & Health Sc.* (39.4%), *Engineering and Technology* (23.6%), *Social Sc.* (6.3%), *Agricultural Sc.* (3.3%) and *Humanities* (1.5%). (Greek Scientific Publications, 1996-2010).

The majority of Greek scientific publications were produced by Universities, Research Centers supervised by GSRT, Public and Private Health Institutions, Technological Educational Institutes, Other Public Research Centers and Other Public and Private Institutions. (Greek Scientific Publications, 1996-2010)

Data shows the wide use of English (99,66%) in Greek publications, as national and international collaboration of the scientific community allows the authors to increase the visibility, number of citations and the impact of their publications. Other common languages used: French (0.110%), German (0.103%), Spanish (0.032%) and Italian (0.014%).

Table 1 indicates the research collaboration between Greece and other countries. The evolution over the period 1996-2010 showed an increasing trend in a national and international level. In 2010 co-publications by Greek researchers accounted for 67.2% of the total publications output (49.3% in 1996). During 2006-2010 Greece cooperated with scientists from 154 countries. The main publishing partners were USA, UK, Germany, France and Italy. (Greek Scientific Publications, 1996 -2010)

Table 1: Scientific collaboration between Greece and other countries

Countries/Territories	Record count	% of 65.237
GREECE	65237	100.000%
USA	7601	11.651%
ENGLAND	5518	8.458%
GERMANY	4557	6.985%
FRANCE	3552	5.445%
ITALY	3347	5.131%

During 1996-2010, Greece exhibits a remarkable growth rate in its publishing volumes and is ranked 8th among the OECD-34 countries. Featuring 10,219 publications in 2010 compared to 3,729 in 1996, Greece presents a rate of change equal to 2.74, above the average rate of change for the EU (1.54) and the OECD countries (1.41) The number of Greek publications displayed a steady increase from 1996 to 2008. However, this positive trend reversed in 2009; the rate of change in Greek publications was

almost zero that year. The situation deteriorates in 2010 with a decline higher than that observed in EU and OECD countries.

Factors affecting the growth and the recession of Greece's publication profile

Greece's entry to the European Union (EU) had a very positive impact on the development of the research component of Greek universities mainly, but also on the sustainability and development of Greek Research Centers (RCs). Significant inflows from structural and competitive programs of the Union oxygenated research. Meanwhile, a number of structural changes in the academic legislation radically changed positively the landscape of higher education: In 1983 (Law 1404/83) Technological Educational Institutes (TEI) were founded and in 1984 three new universities: *University of the Aegean*, based in Mytilene, *Ionian* based in Corfu and *Thessaly* based in Volos (Presidential Decree (83/84) 31/A/20-03-1984). In 1985 (Law 1514/85) the National Council for Research and Technology (ESET) along with the Foundation for Research and Technology (FORTH) in Crete, the Academic RCs and Postgraduate Institutes were established.

Period 1985-1992 and the decade that follows are characterized by dense flow of resources through the structural and competitive programs in the EU. In 1992 the Greek Open University was founded.

The main sources of funding for research in Greece are: Grants from regular state budgets, From the public investment program of GSRT (it includes EU Structural Funds (eg. NSRF) for 2007–2013 and constitutes the reference document for the programming of EU Funds at national level) and Funds from other international collaborative programs (NSRF, 2007-2013: *What is the National Strategic Reference Framework*).

In recent years reports of abuses, underfunding, economic mismanagement, interventions in the evaluation of research programs and delays by the Ministry of Education, affected negatively Greece's research productivity in publications. Voices of academics proliferate, complaining to the Ministry of Education in order to expedite the process of evaluating research programs of the NSRF, or precious funds from the EU will be lost (Federation of Traning Personal of TEI, 2011).

Today, Greece is going through difficult times in the context of the economic crisis that plagues the whole of Europe. The great achievements of Greek researchers in recent years seem to spend a period of recession. However, ″*Small countries with no oil or diamonds, have the power of human capital. Our strength is the heads* [of researchers]″ (Diamantopoulou A, 2011)

References

National Documentation Centre of Greece, (2012). [online]. Retrieved September 20, 2012 from: www.ekt.gr/content/display?ses_mode=rnd&ses_lang=el&prnbr=84879>

Greek Scientific Publications, (1996-2010. Retrieved 20/09/2012 from: http://metrics.ekt.gr/en/report02/index

Diamantopoulou, A. (2011). Extract interview on NET TV, 27/5/2011. Retrieved 30/09/2012 from: http://www.diamantopoulou.gr

Jeenah, M., & Pouris, A. (2008). S. African research in the context of Africa and globally. *S. African*

Journ. of Science, 104 (9/10), 351-354
King, D. A. (2004). The scientific impact of nations: What different countries get for their research spending. *Nature, 430*, 311–316
Pouris, A. [1](2012). Science in S. Africa: The dawn of a renaissance? *S Afr J Sci.* 108(7/8): 66-71
Pouris, A. [2](2012). Scientometric research in S.A. and successful policy instruments. *Scientometrics* 91: 317–325
Pouris, A. (2003). S. As publication record: the last ten years. *S. A. Journal of Science 99,* 425-428
NSRF, (2007-2013): *What is the National Strategic Reference Framework* [online] Retrieved 20/09/2012 from: http://www.espa.gr/en/Pages/staticWhatIsESPA.aspx
FEDERATION of TRAINING PERSONAL of TEI :press release, 07/01/2011. Retrieved 4/09/2012 from: http://www.teiath.gr/verwalt/syllogoi_foreis/osep_tei/

THE ROLE OF COGNITIVE DISTINCTIVENESS ON CO-AUTHOR SELECTION AND THE INFLUENCE OF CO-AUTHORING ON COGNITIVE STRUCTURE: A MULTI-AGENT SIMULATION APPROACH

Bulent Ozel[1] and Ahmet Suerdem[2]

[1] *bulent.ozel@bilgi.edu.tr*
Istanbul Bilgi University, Dolapdere Kampusu, Beyoglu, 34440 Istanbul
(Turkey)

[1] *ahmet.suerdem@bilgi.edu.tr*
Istanbul Bilgi University, Kustepe Kampusu, , 34440 Istanbul
(Turkey)

Motivation

Analysis of co-authorship relations is a tangible and reliable way of tracking scientific collaboration networks. Particularly, they give important information about knowledge diffusion. Habitually, bibliometric network analyses focus on the effects of author attributes on author interactions such as homophily or compositional measures. However, the interaction between collaboration networks and cognitive structures is a relatively less studied area. Particularly, how co-authors may influence each other's knowledge structures and the role of cognitive homophily or heterogeneity for the selection of potential co-authors is essential to understand knowledge diffusion within an academic domain. Studies up to now show how author attributes affect whom to collaborate while underplaying how this interaction influences the cognitive structures of the collaborating authors. Our objective is to analyse not only how collaboration network changes as a function of itself and author attributes but also how the author cognitive structures change as a function of themselves and of the network.

Methodology

We use a stochastic actor-based model to test the effects of cognitive structures on network change and the effects of network change on cognitive structures. Our model is an actor driven model where each author is capable of selecting its co-author and his/her interest area. We operationalize cognitive structures of individual actors as semantic networks. Cognitive similarities are calculated according to the co-occurrence of the subject keywords within the articles. We test the hypothesis that authors would choose to collaborate with authors cognitively distinct from them since there would be more possibilities for cross-fertilization compared to cognitive homophily.

Findings

Results from our initial experiments hint that in scenarios where agents are inclined to collaborate with cognitively

dissimilar agents, then resulting collaboration structure rather mimics co-authorship relations seen within a research center. On the other hand, when cognitive similarity leads the incentives to pick a collaborator, then resulting co-authorship rather mimics network structures observed within domain of a journal in a field.

A large set of experiments are to be conducted to fully verify and validate our initial results as well as to discuss challenges addressed above. There are a set of additional implementation challenges, which will be addressed and attempted. They are (i) how to model when and in what circumstances multiple co-authorship occurs; (ii) and how to specify knowledge content of collaboration. Cognitive structures of interacting dyads will be studied while addressing these questions. Besides, (iii) at each run, not only new knowledge pieces but also new agents will be injected to the simulation. Knowledge base of those new agents will be composed of partially by a subset of keywords that is already in the current set and partially by new keywords that is not in the set. This approach will mimic arrival of new scientists in a field.

SCIENTIFIC PRODUCTION AND INTERNATIONAL COLLABORATION ON SOLAR ENERGY IN SPAIN AND GERMANY (1995-2009)

Elías Sanz-Casado[1]; Maria-Luisa Lascurain-Sánchez[1]; J. Carlos García-Zorita[1]; Antonio Eleazar Serrano-López[1]; Birger Larsen [2] and Peter Ingwersen [1,2,3]

[1] elias@bib.uc3m.es; mlascura@bib.uc3m.es; czorita@bib.uc3m.es; aeserran@bib.uc3m.es; [3]blar@iva.dk : [2,3] PI@iva.dk

[1]Universidad Carlos III de Madrid, Department of Library Science and Documentation, Laboratorio de Estudios Metricos de Informacion (LEMI), C/Madrid 126, Getafe 28903, Madrid, Spain.

[2]Royal School of Library and Information Science, Biketinget 6, DK 2300 Copenhagen,S, Denmark

[3]Oslo University College, St. Olavs plass, 0130 Oslo, Norway

Introduction

Renewable energies carry political and financial significance in all EU countries. Their importance is translated into a major research and innovation trend, particularly in relation to the achievement of sustainable resources (Walz; Schleich & Ragwitz, 2011).

Solar and Wind Energies offer the biggest potential for energy production, as it has been highlighted in the last decade (Sanz-Casado; García-Zorita; Serrano-López; Larsen & Ingwersen, 2012). Within he overall conglomerate of renewable energies, Germany has a bigger production than Spain, although the increase is higher for Spain in the case of Solar Energy production (2560 versus 2734 of increment), measured in tonnes of oil equivalent (during 1995-2009).

Objectives

The overall objective of this research is to identify if the public interest in Solar energy is reflected in Spanish and German research publications, so that we can establish a profile in the evolution of publications and in their international scientific collaboration patterns overtime.

Methodology

We have used the databases contained within the Web of Science as source. A search strategy was designed to be able to retrieve publications on Solar energy from a wider collection of documents on renewable energies. The timeframe goes from 1995 to 2009, while the geographical framework focused on documents by authors from Spanish and German institutions.

The obtained records were downloaded onto a text file, using a set of scripts with Perl programming language for their treatment in a referential database managed by MySQL.

Results

Table 2 shows the distribution of scientific production on Solar Energy by country (Top-10) (WoS, 2012), with a remarkable raise of China to second position and a decrease of USA and EU countries who produce more articles but represent a lower percent over the global

production on Solar energy against that was found on Wind power research patterns (Sanz-Casado; García-Zorita; Serrano-López; Larsen & Ingwersen, 2012). For instance, DEU gains in publications but decreases in percentage. The same occurs for France but not for Italy that increases its percentage. Spain increases both in number of publications and in percentage (to 4.14 from 2.89).

Table 1. Supply, Transformation and Consumption - Renewable and Wastes (total, solar heat, biomass, geothermal, wastes) - Annual Data. (EUROSTAT 2012)

Renewable Energies			
YEAR	**EU**	**Germany**	**Spain**
1995	82,631	6,095	5,510
1996	86,139	6,278	6,986
1997	89,754	7,228	6,646
1998	92,353	7,795	6,783
1999	92,681	8,069	6,031
2000	96,650	9,094	6,928
2001	99,637	9,747	8,169
2002	97,505	10,898	7,040
2003	103,906	12,969	9,245
2004	111,843	15,780	8,866
2005	115,891	17,502	8,353
2006	123,507	21,678	9,158
2007	134,057	27,964	9,996
2008	142,037	27,968	10,334
2009	148,776	27,777	12,158
Solar Energy			
YEAR	**EU**	**Germany**	**Spain**
1995	282 (0.34%)	38 (0.62%)	26 (0.47%)
1996	305 (0.35%)	48 (0.76%)	26 (0.37%)
1997	329 (0.37%)	61 (0.84%)	24 (0.36%)
1998	362 (0.39%)	77 (0.99%)	27 (0.4%)
1999	391 (0.42%)	92 (1.14%)	29 (0.48%)
2000	430 (0.44%)	115 (1.26%)	33 (0.48%)
2001	482 (0.48%)	150 (1.54%)	38 (0.47%)
2002	533 (0.55%)	184 (1.69%)	43 (0.61%)
2003	594 (0.57%)	216 (1.67%)	48 (0.52%)
2004	683 (0.61%)	262 (1.66%)	58 (0.65%)
2005	806 (0.7%)	353 (2.02%)	65 (0.78%)
2006	988 (0.8%)	472 (2.18%)	83 (0.91%)
2007	1264 (0.94%)	580 (2.07%)	137 (1.37%)
2008	1,730 (1,22%)	735 (2.63%)	352 (3.41%)
2009	2,498 (1,68%)	973 (3,5%)	711 (5,85%)

Table 2. Top-10 Countries Producing Research on Solar Energy- 1995/2009 (WoS, 2012)

1995-1999		**2000-2004**		**2005-2009**	
	Publi-cations		**Publi-cations**		**Publi-cations**
USA	2,580 (28.63%)	USA	3,096 (23.79%)	USA	6,038 (22.25%)
DEU	1,006 (11.16%)	**DEU**	1,669 (12.83%)	CHN	3,535 (13.03%)
JPN	849 (9.42%)	JPN	1,650 (12.68%)	JPN	2,643 (9.74%)
IND	484 (5.37%)	CHN	643 (4.94%)	**DEU**	2,500 (9.21%)
ENG	450 (4.99%)	ENG	612 (4.7%)	KOR	1,241 (4.57%)
FRA	420 (4.66%)	FRA	579 (4.45%)	IND	1,217 (4.48%)
AUS	343 (3.81%)	IND	510 (3.92%)	**SPA**	1,123 (4.14%)
SPA	260 (2.89%)	**SPA**	505 (3.88%)	FRA	1,080 (3.98%)
CAN	259 (2.87%)	NLD	464 (3.57%)	ENG	1,031 (3.8%)
ITA	249 (2.76%)	AUS	445 (3.42%)	ITA	828 (3.05%)
129 countries		131 countries		144 countries	
9,011 docs in total		13,012 docs in total		27,138 docs in total	

International Scientific Collaboration

Figures 1 and 2 show the international research collaboration patterns for both countries with regard to the percentage of publications that present this collaboration, as well as the number of authors from different institutions by country.

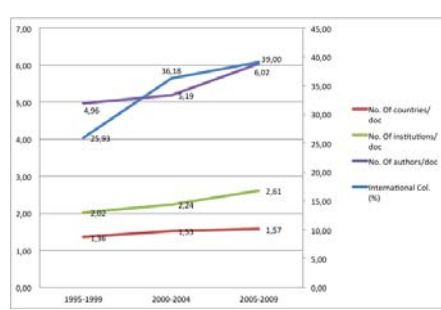


Figure 1. Collaboration Patterns in Solar Energy (Germany). WoS, 2012

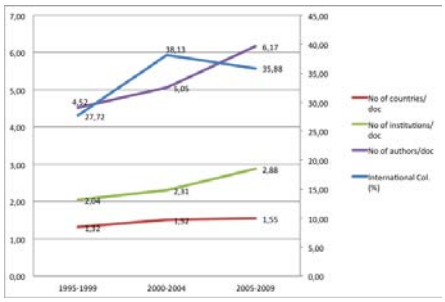


Figure 2. Collaboration Patterns in Solar Energy (Spain). WoS, 2012

Table 3 shows the countries that collaborate in Spain and Germany publications. For Spanish researchers, Germany is the first country with they collaborate, then USA. USA is the main partner of Germany in this area, while Spain ranks fourth.

Table 3. Top-10 countries collaborating with Spain and Germany on Solar Energy research -1995/2009 (WoS, 2012).

SPAIN		GERMANY	
No. of Countries 61	**No. of Docs. 1,716**	**No. of Countries 81**	**No. of Docs. 4,793**
Country	Docs.	Country	Docs.
DEU	163	USA	369
USA	97	GBR	176
GBR	77	FRA	171
FRA	76	SPA	163
CHE	43	CHE	141
ITA	42	NLD	139
PRT	31	JPN	121
MEX	28	ITA	98
ISR	24	AUS	97
NLD	24	AUT	88

Conclusions

Germany represents the top EU countries in the three different periods analysed, although its presence decreases in percentages. The case for Spain is the opposite, as the percentage of publications increases during the third period and it goes up one position to be among the Top-10 countries.

The collaboration profiles for both countries are very similar, despite the fact that the percentage of international collaboration for Spain decreases during the third period.

Germany is the top country regarding collaboration with Spain, followed by the USA, while the USA is the top country regarding collaboration with Germany, with Spain coming into fourth place in this context.

Acknowledgments

This research was funded by the Spanish Ministry of Economy and Competitiveness under the project CSO2010-21759-C02-01 entitled, "Análisis de las capacidades científicas y tecnológicas de la eco-economía en España a partir de indicadores cuantitativos y cualitativos de I+D+i", carried out by the Carlos III University of Madrid.

References

Sanz-Casado, E., García-Zorita, J. C., Serrano-López, A. E., Larsen, B., & Ingwersen, P. (2012). Renewable energy research 1995–2009: a case study of wind power research in EU, Spain, Germany and Denmark. *Scientometrics,* 95 (1), pp 197-224.

Walz, R., Schleich, J., & Ragwitz, M. (2011). Regulation, Innovation and Wind Power Technologies–An empirical analysis for OECD countries. *Paper presented at the DIME Final Conference.*

Web of Science. Available at: [http://thomsonreuters.com/products_services/science/science_products/a-z/web_of_science/].

SCIENTIFIC PRODUCTION OF TOP BRAZILIAN RESEARCHERS IN BIOCHEMISTRY, PHYSIOLOGY, PHARMACOLOGY AND BIOPHYSICS

Daniel Henrique Roos; Luciana Calabró; João Batista Teixeira da Rocha; Diogo O. G. de Souza.

daniel_varzeano@hotmail.com; luciana.calabro.berti@gmail.com; diogo@ufrgs.br; jbtrocha@yahoo.com.br
Federal University of Rio Grande do Sul, Departamento de Bioquímica, Rua Ramiro Barcelos, 2600 – anexo, Porto Alegre, RS (Brazil)

Introduction

In many developing countries, including Brazil, there has been an increasing awareness of the need to develop science. In Brazil institutionalization of science is recent, when compared to Europe and USA [1]. In recent decades, an expressive expansion of scientific production in Brazil has occurred in biological sciences. Currently, the country accounts for 46.6% of the scientific production in Latin America and 1.75% of world production. About two decades ago the country accounted for 0.5% of world production [2]. Main reasons for this increase are the stability of the investment in research and changes in the policies of the main funding agencies [1].

The National Council for Scientific and Technological Development (CNPq) supports monetarily researchers with high scientific productivity. The financial support is for particular use (all grades of researchers, i.e., PQ-2 (the lowest level), PQ 1D, 1C, 1B and 1A (the highest) and for the use in the lab (for level 1 researchers). There is also a financial support to senior researchers (normally retired ones) to be used in research related activities (PQ-Sr). This work aims determine and compare the profile of publication in scientific journals of researchers in the level 1 (PQ 1C-A) in order to get a picture of the publications by the top investigators in biochemistry, pharmacology, biophysic, physiology and basic neuroscience (which is called BF in CNPq and comprises the biological sciences II in the Coordination for postgraduate personnel improvement, CAPES). The results presented here can inform CNPq and CAPES about the productivity of PQ-1 of Biological Sciences II and thus partially contribute to determine the effectiveness of funding science in Brazil.

Methodology

The present study analyzed the scientific production of senior researchers' fellows from CNPq in the subareas of Physiology (including basic neurosciences), Biochemistry, Pharmacology and Biophysic (which is called BF). The scientific production of researchers was analyzed during all productive live of researchers.

The data were obtained for all researchers' fellows included in the CNPq category of PQ-1 (1A, 1B, 1C). The total number of PQ-1 in BF area of

CNPq is 228. The analysis included: number of articles published; authoring type (publication as first author or as last author), number of citations; *h*-index. These data were obtained from the database of Scival - Scopus).

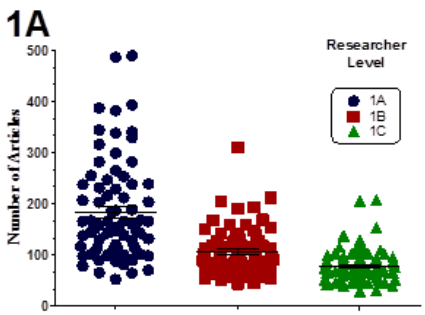


Fig. 1 Number of article published by researcher fellows; circle 1A; square 1B; triangle 1C; n= 76.

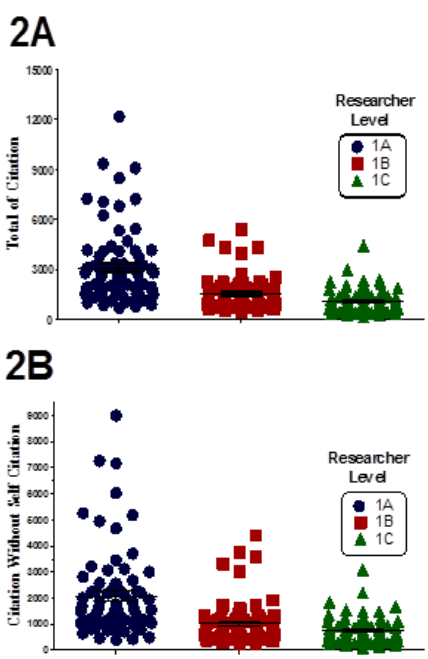


Fig. 2 Number of citation by researcher 2A; number of citation by researcher without self-citations of all authors of the article 2B; circle 1A; square 1B; triangle 1C; n= 76.

Results

As expected the number of articles published by researcher fellow decreased gradually with researcher level (1A, 1B, 1C respectively; Fig. 1). The results also demonstrated that the some of these researches have published more than 300 articles in journals evaluated by scopus (www.scopus.com) during their career .

The average of citations by researchers group is shown in Fig. 2A. In order to analyze the influence of articles in the literature, the data were evaluated without the self citations of all authors (Fig. 2B). The results indicated that the number of citation decrease ≈ 32% when self citations of all authors were excluded (Fig 2 A and B). Similarly, the *h*-index of researchers decrease ≈ 23% when self citations were excluded.

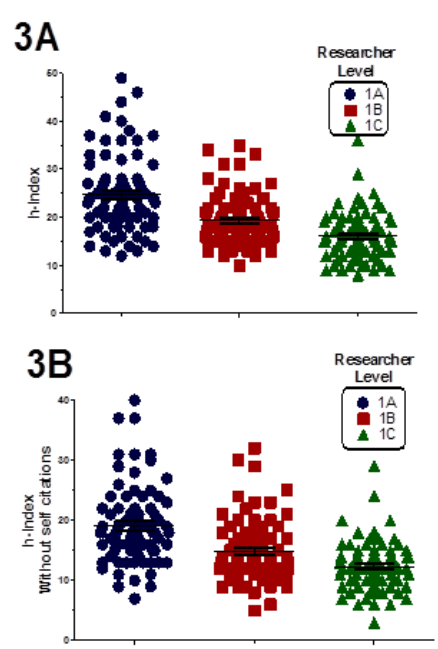


Fig. 3 h-Index of researchers 3A; h-Index of researchers without the self-citations of all authors of the article 3B; circle 1A; square 1B; triangle 1C; n= 76.

The Figure 4 exhibit the number of articles as first and last author published in the last 10 years. This figure shows the autonomy of research, since the last/first author is expected to answer questions of reviewers and interacts with the journal and the scientific community.

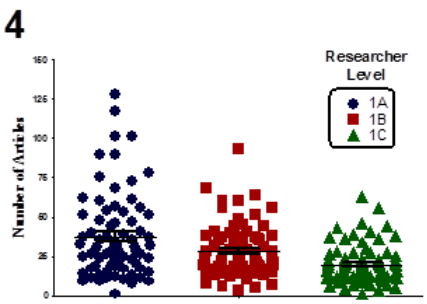


Fig. 4 The number of articles by researcher fellows as first and last author published in last 10 years; circle 1A; square 1B; triangle 1C; n= 76.

Conclusions

The performance of top Brazilian researchers (PQ-1) in the BF area of CNPq reveals that there no was homogeneity in the quantitative/qualitative parameters evaluated within the sublevels 1A, 1B, 1C. However, it is possible to observe that the mean of quantitative parameters analyzed varies in accordance with the researcher level. However, the reasons for the overlapping between the research levels deserve a more detailed analysis.

Acknowledgements

This work was supported by grants from, UFSM; UFRGS; FAPERGS; CAPES; CNPq; FINEP (IBN-Net) and INCT-EM.

References

[1] Zorzetto R., Razzouk D., Dubugras M.T.B., Gerolin J., Schor N., Guimarães J.A. Mari J.J. (2006) The scientific production in health and biological sciences of the top 20 Brazilian universities. Br. J. Medic and Biol. Res. 39: 1513-20

[2] De Meis, L. Arruda, A.P Guimarães, J. (2007). The Impact of Science in Brazil. *IUBM Life*, *59*(4), 227-234.

A SIMPLE METHOD TO ASSESS THE QUALITY OF ANY UNIFICATION PROCESS

Mendez-Vasquez Raul Isaac[1]*; Vila Domènech Joan Salvador[2]; Suñén-Pinyol Eduard[3]*; Olivé Vázquez Gerbert[4]

[1]*rmendez@prbb.org and* [3]*esunen@prbb.org*
Bibliometrics Unit, Barcelona Biomedical Research Park Foundation (FPRBB). C/ Doctor Aiguader, 88 E08003 Barcelona, (Spain)

[2]*jvila@imim.es and* [4]*jolive@imim.es*
[2]Statistical Unit, [4]Computing resources, IMIM-Institut Hospital del Mar d'Investigacions Mèdiques. C/ Doctor Aiguader, 88 E08003 Barcelona, (Spain)

**Present address:* [1]*raul.mendez@fundaciorecerca.cat and* [3]*eduard.sunen@fundaciorecerca.cat*
Fundació Catalana per a la Recerca i la Innovació (FCRI). Psg. Lluís Companys, 23 E08010 Barcelona, (Spain)

Introduction

According to Van Raan (2005) the attribution of publications to organizations is one of the most important technical problems to solve when building reliable bibliometric reports.

Attributing publications to a specific organization based on the data in address field is not trivial, and doing it right may be problematic and time consuming, especially when the focus of studies moves below the national (Butler 1999).

The presence of variants of names of organizations, and of locations, in citation indexes has been reported as one of the main sources of troubles (Anderson 1998), (Leydesdorff 1988), and (Bourke1996). Added, a number of situations increase the complexity of this process: 1) structural and name changes in organizations (Van Raan 2005), 2) multi-located institutions (de Bruin 1990), 3) presence of multi-affiliated researchers (Butler 1999), 4) missing pieces of information about "mother" organizations in addresses (de Bruin 1990), or 5) interactions with other institutions. Further, these factors may also show regional peculiarities.

On the one hand, it is very likely that authors will continue to enlarge the list of variants of names and locations in citation indexes. On the other, acknowledging that there is certain level of error in every measure might be of help in finding better ways to deal with this problem. From this perspective the main challenge in producing reliable reports would be achieving a reasonably high level of precision (admitting that error will always be present) in a reasonably short period of time which may be more interesting to final users.

The aim of the present study is to show a simple method to ensure an "at most" percentage of error in any unification process.

Methods

Modeling the problem. The number of addresses correctly mapped to

respective organizations can be used as indicator of the quality of an unification process. Thus, assessing its quality would only require examining addresses to classify each case as either "correctly" or "wrongly" attributed. Such type of experiments are Bernoulli trials, as only one of the two outcomes is possible, and a series of experiments have a binomial distribution X~*B(n,p)*.
Apparently the binomial function would solve our challenge. However, examining a fixed number of addresses could lead to the analysis of non-representative samples of address, and examining all addresses is simply not feasible because their large number. Fortunately, a series of independent Bernoulli trials also have negative binomial distributions X~*NB(r,p)*, which provide the probability distribution of the number of successes before a specified number of failures "r" occur. In our case a success corresponds to a wrongly attributed address and failures to the number of correctly attributed. The probability of success *p* is 0.03 in the case of a maximum percentage of error of 3% and 0,05 for a maximum of 5%.

Results

Implementing the solution. Using a NB enable setting in advance an upper threshold to 1) the percentage of error that we "tolerate" *p*, and 2) the statistical confidence of the test. In Table 1 the sizes of the samples and the maximum number of errors are calculated using the *pnbinom()* function in R for maximal levels of error of 3% and 5%, with 95% and 99% statistical confidence respectively. Thus for example, assuming that addresses have been attributed with a 3% error, there are 95% chances that a wrong case occurs in a random sample of 99 addresses. So, if the examination delivers no wrong case, we can ensure, with a 95% confidence, that the error is less than 3%.
The result of performing theses phases iteratively (sampling, examination and correction) is a quality assessment (QA) method for achieving a particular level of accuracy in any unification process.

Discussion

The presence of variants of names and locations in citation indexes causes troubles during the unification of addresses, a problem known for quite long time now (Butler 1999), (de Bruin 1990).

Table 1. Sizes of the random samples and maximum number of errors tolerated

95% conf. error < 3%		**95% conf. error < 5%**	
#trials[1]	#errors[2]	#trials[1]	#errors[2]
99	0	59	0
157	1	93	1
208	2	124	2
257	3	153	3
303	4	181	4
348	5	208	5
99% conf. error < 3%		**99% conf. error < 5%**	
#trials[1]	#errors[2]	#trials[1]	#errors[2]
152	0	90	0
219	1	130	1
277	2	165	2
332	3	198	3
383	4	229	4
433	5	259	5

1, size of the random sample of addresses; 2, maximum number of errors ("wrongly" attributed addresses) allowed in the sample.

Surprisingly, to our knowledge there are no publications describing methods to assess the accuracy of this process.
The present method allows ensuring, by statistical means, a specific level of accuracy of any unification process when applied iteratively. The benefits of its application include: 1) ensuring a specific level of accuracy independently of the level of the study (*macro*, *meso* or

micro), 2) enables assessing the quality of unifications of locations, 3) enables aligning "delivery time" and client needs, since early previews or "reports on trends" may be provided based on data unified at different levels of accuracy, 4) enables identifying "big" errors caused by recent changes in the structure of organizations and also rare variants, both of which improves the efficiency and the quality of the unification.

The results of the application of the present method will depend on the definition of the "wrong case", which may vary on a study base.

In our opinion, a less than pleasant scenario lies ahead for our community in the next years. It is likely that errors in addresses will continue to occur even in greater number in citation indexes, and the "*press the button*" attitude is very likely to spread even more among final users of bibliometric reports. The combination of these trends will increase the pressure on practitioners committed to working with scientific rigor. We hope that this method provides a new perspective for dealing with errors.

Limitations

The present method does not address the issue of missing addresses; it is based on existing data. As for not fully informative address, they are equally probable of being sampled during the QA. It depends on both, reviewers to unify them, and on experts to tag them as "correct" or "wrongly" attributed according to the aim of studies. Future studies should address the issue of missing addresses in order to estimate the size and the effect of this phenomenon on bibliometrics indicators.

Conclusion

The present method provides a new approach to producing reliable reports with a reasonably high level of precision in a reasonably short period of time.

Measuring the magnitude of "error" and eventually its effect on indicators, instead of neglecting its presence in reports, is a scientific attitude that will probably benefit all practitioners. Hopefully this approach will also lead to new ways of interpreting bibliometric reports more cautiously and responsibly, which in turn will probably improve the acceptance of bibliometric methods inside and outside our community. However, a widely agreed-upon recommendation for best practice in normalizing addresses is absolutely necessary to improve comparability of bibliometric reports.

References

Anderson J., Collins, P. M. D., Irvine, J., Isard, P. A., Martin, B. R., Narin, B. R., Stevens, K. (1988). On-line approaches to measuring national scientific output: A cautionary tale, Science and Public Policy, 15,153–161.

Bourke, P., Butler, L. (1996). Standards issues in a national bibliometric database: The Australian case. Scientometrics, 35,199–207.

Butler L. (1999). 'Who 'Owns' this Publication?', in Proceedings of the Seventh Conference of the International Society for Scientometrics and Informetrics (ISSI '99) (pp. 87-96).

de Bruin R.E., Moed H.F. (1990). The unification of addresses in scientific publications. In: Egghe L., Rousseau R. (editors.), Informetrics, (pp 65-78) Elsevier Science Publichers B.V.

Leydesdorff, L. (1988). Problems with the 'measurement' of national

scientific performance. Science and Public Policy, 15,149–152.
Van Raan A. F. J. (2005). Fatal attraction: Conceptual and methodological problems in the ranking of universities by bibliometric methods. Scientometrics, 62(1),133-143

STRUCTURE ANALYSIS OF SMALL PATENT CITATION NETWORK AND MAPPING TECHNOLOGICAL TRAJECTORIES

Yanling Wang[1,2]

violin@whu.edu.cn

[1] Information Management School, Wuhan University, 430072, Hubei Province, P.R.China

[2] Wuhan Institute for Science of Science, 430023, Hubei Province, P.R.China

Introduction

The theory of technological trajectories argues that the development of a technology is along some specific technology trajectories. The technology trajectories can be used to explore the history of a technology area or to predict the technological future or discovery a new technology opportunities. Past researches on mapping technologies as patent citation networks mostly based on the large network, such as the fuel cell patent citation network (Verspagen 2007)which is composed of 15506 nodes and more than one million relations. The patent citation network of coronary artery disease was composed of 5,136 patent documents that granted between 1979 and 2003 (Mina, Ramlogan et al. 2007). There are 22,095 nodes in the network.

However, as we know some patent networks of specific technology area are not as large as those networks I mentioned before. There are maybe only hundreds of nodes in the network, and the scale of the citation matrix mostly is not more than 100,000. If the research locates at these technology areas, the method to discover or identify the trajectories is not as same as that applied in large networks.

Radical technological breakthroughs and incremental technological innovations are considered as the different technology evolution paradigm. No matter radical or incremental technology innovations are specific technological trajectories. And the critical technology (or patent) just like a chain of pearls, composes the track of the development of a technology. Technological trajectory theory argued that there must be a method to identify the trajectory no matter the area is broad or just narrow. This paper explored the method to discover the key technology nodes or technology trajectories of narrow area. With the thorough analysis of the structures of two small patent citation networks, the characteristics of a small network have been concluded.

Data

Patent data this paper used come from the BASICBIB database and USCITES database (Bhaven, 2011), which were constructed by Bhaven N. Sampat professor from Columbia University. The database BASICBIB includes basic "front page" data for patents issued from January 1, 1975 to December 31, 2010. The database USCITES includes U.S. patent citations in utility patents issued from1/1/1975 to 12/31/2010. Each observation is a citing-cited pair. These two database are based mainly on information from a custom extract DVD generated by the Electronic Information

Products Division of the USPTO on 4/29/2011. Two of the variables (application number and application date) were extracted from the USPTO's CassisBIB Patents DVD.

We use the BASICBIB and USCITES database to map patent citation network of photovoltaic technology. The USPC class of the field of photovoltaic technology is defined in the USPTO EST Concordance186, and the USPC is "136/243+". We searched in BASICBIB, there are 36 patents of this field. We search the patent citation data in USCITES joined with BASICBIB, there are 462 patent backward citations during 1979-2010. We use UCINET to construct patent citation network of 482*482.

Methods

Hummon and Doreian(Hummon 1989) assign a weight to each citation link on the basis of its position in the overall structure of the network. The method is based on the examination of the different 'search paths' existing in the network. Search paths are sequences of links that connect the vertices of the network.

Main Paths Analysis

The Main Path algorithm identifies the most important papers and streams of growth or development in a citation network.[187] (Critical Path Method, CPM)

"The main paths should be viewed as the main flow of ideas characterizing the structure of the network in question."(Roberto Fontana 2008)

Subnetwork Extraction

Components of a graph are sub-graphs that are connected within, but disconnected between sub-graphs. In addition to identifying the members of the components, UCINET calculates a number of statistical measures of graph fragmentation.

Table1 Index of Network Structure of Photovoltaic Field (1979-2010)

Period	*Density*	*No. of Ties*	*Components*	*Normalized heterogeneity*	*Entropy*	*Distance-based cohesion ("Compactness")*
1979-1995	0.0619	13	2	0.248	0.145	0.062
1979-2002	0.0244	29	6	0.795	0.458	0.024
1979-2003	0.0117	70	8	0.803	0.409	0.012
1979-2008	0.0064	129	13	0.893	0.471	0.006
1979-2009	0.0030	308	15	0.696	0.316	0.003
1979-2010	0.0020	462	25	0.765	0.354	0.002

Results and Discussion

In order to find the reason of the failure, we conclude the index of the network structure of different period, which is illustrated in Table 1. All the index of network structure is calculated by UCINET.

[186]The environmentally sound technologies (EST) concordance was issued by USPTO in 2010. http://www.uspto.gov/web/patents/classification/international/est_concordance.htm

[187] Hummon and Doreian (1989) devised three indices or weights ofedges to computationally identify the (most) important part of a citationnetwork—its main path. Batagelj (2002) developed algorithms toefficiently compute the Hummon and Doreian's indices so that they canbe used for the analysis of very large citation networks with severalthousands of vertices.

Although the network grows up gradually and the nodes in the net become more and more, the whole net is very loose. The characteristics of small patent citation network:low density, high Fragmentation, more and more components, short distance.

The network of developer field is also characterized by the four characteristics. Figure 9-Figure13 are the patent maps of networks of developer field in different period. Table 1 is the index of network structure of developer field.

From the analysis, we find that the low density, high fragmentation, scattered components and short distance cause the methods to extract main path of Hummon and Doreian fail. However, we can search the nodes with high degree centrality or betweeness centrality. These nodes which represent the patents that have dense connectivity with other patents are the main technology in the field. Table 2 is the list nodes with high centrality of different network.

We select the index of network structure: Network Centralization, Heterogeneity, Normalized Heterogeneity, mean of degree and standard deviation to analyze the structure of the network of components. From the patent maps of the components, only the figure 16 and figure 17 are completely centered, and from the patent in the core to the other patents, there are at least 20 or 30 ways, this kind of network only have one step of path. We defined this kind of network as unable to be extracted a path from it. The component can be used to represent the trajectory of technology is characterized:

Low network centralization. The centralization of the network cannot be too high. If the centralization is very high, the network always has star-like structure. There is only one way from one patent to the patent in the core, and other patent is isolated with each other. For this kind of network, the main path is equal with the net and in other words that there is no main path or trajectory in this kind of network. In figure 14 and figure 15, the centralization of network is 60.6%.In figure 18 and figure 19, the centralization of network is 84.86% and 98.22%, and the longest path in those networks is 4-step or 5-step. In figure 23, the centralization of network is 33.31%, and the longest path is 13-step. And in other figures of patent map of main components, the path is only one-step.

Long path between start node to end node. As we analyzed in the network centralization. The structure of network is not star-like, but long and narrow with long path.

Low Heterogeneity. The heterogeneity of the network that figure23 illustrate is 5.92%, and figure 14 and figure 15 is 16.32%, figure 18 is 19.11%, figure 19 is no more than 25%.

Conclusion

The past researches on mapping technologies as patent citation networks mostly based on the large network, This paper explored the method to discover the key technology nodes or technology trajectories of narrow area. With the thorough analysis of the structures of two small patent citation networks, the characteristics of a small patent citation network and main component of those nets have been concluded.

The characteristics of small patent citation network: (1) Low density; (2) High Fragmentation; (3) More and more components; (4) Short distance.

The component can be used to represent the trajectory of technology is characterized: (1) Low network

centralization.; (2) Long path between start node to end node; (3) Low Heterogeneity.

References

Batagelj, V., Mrvar, A., 2003. Pajek: analysis and visualization oflarge networks'. Department of Theoretical Computer Science,University of Ljubljana, Preprint series, Vol. 41, p. 871.

BHAVEN, "USPTO Patent and Citation Data", http://hdl.handle.net/1902.1/16412 UNF:5:ERqPZ7enbwBRimghqDD4 gQ== Bhaven Sampat [Distributor] V4 [Version]

Mina, A., R. Ramlogan, G. Tampubolon and J. S. Metcalfe (2007). "Mapping evolutionary trajectories: Applications to the growth and transformation of medical knowledge." Research Policy 36(5): 789-806.

Roberto Fontana, A. N., Bart Verspagen (2008). "Mapping Technological Trajectories as Patent Citation Networks. AN Appllication to Data Communication Standards." SPRU Electronic Working Paper Series No. 166.

Verspagen, B. (2007). "Mapping technological trajectories as patent citation networks: A study on the history of fuel cell research." Advances in Complex Systems 10(1): 93-115.

Borgatti, S.P., M.G. Everett, and L.C. Freeman. 1999. UCINET 6.0 Version 1.00. Natick:Analytic Technologies.

STRUCTURE OF INTERDISCIPLINARY RESEARCH: COMPARING LM AND LDA

Qi Wang[1] and Ulf Sandström[2]

[1]*qiwang@kth.se*
INDEK, KTH-Royal Institute of Technology, Lindstedtsvägen 30
100 44 Stockholm, Stockholm (Sweden)

[2]*ulf.sandstrom@indek.kth.se*
INDEK, KTH-Royal Institute of Technology, Lindstedtsvägen 30
100 44 Stockholm, Stockholm (Sweden)

Introduction

Interdisciplinary research (IDR) is becoming increasingly important in current academic research. It is widely recognized that IDR is related to progress, creativity, and innovation. Building upon recent literatures, discussions on IDR focus on different topics. That makes it far more challenging to discern the discussions on IDR.

This study intends to solve two questions:

1. From what perspectives current interdisciplinary research are discussed when researchers refer to interdisciplinary research in their articles?
2. Which method is best for detecting similarity?

Methods

Literature data analysed in this study were retrieved from Web of Science. We selected articles that use "interdiscipl*", "multidiscipl*", and "transdiscipl*" as the topic word. The reason we selected multidisciplinary and transdisciplinary as research subject is, in a broad sense, that those three terms, are highly related. They all indicate the integration of "information, data, techniques, tools, perspectives, concepts, and/or theories from two or more disciplines or bodies of specialized knowledge" (Committee on Facilitating Interdisciplinary Research and Committee on Science, 2005). However, the difference among them lies in the degree of integration (Rosenfield, 1992; Wickson et al., 2006). However, we do not intend to discuss this difference, therefore the terminology "interdisciplinary research" is used to represent all research associated with interdisciplinary, multidisciplinary and transdisciplinary activities. In this manner, 47,631 articles including proceeding papers have been downloaded from the year 2002 to 2011. In order to filter noise data, first, we aggregate the articles by Louvain Method (LM), and 1,196 clusters are identified. Through scanning through abstracts of articles, we found that some clusters are not associated with IDR; especially most of the clusters were articles asking for more interdisciplinary teamwork in medicine, but did not contribute to the discussion on IDR in itself. That might hide some categories with small number of articles, and thus our judgment would be affected. To delete those clusters, we analysed abstracts of the two most cited papers in each cluster as we assumed that, the

most cited paper could better represent the category it belongs to. Then, if in our judgment abstracts were not discussing IDR, articles in these categories that were deleted. To make the database as accurate as possible, we repeated the previous step and obtained 2,930 articles finally. Therefore, we have good reason to believe that the left 2,930 articles are associated with interdisciplinary discourses.

LM and Latent Dirichlet Allocation (LDA) are applied to identify topics on IDR. LM is a modularity-based clustering method; in this case, similarity matrix generated from bibliographic coupling (Kessler, 1963) is used as input value. The reason we choose bibliographic coupling is that, previous studies found that bibliographic coupling method generated the most accurate cluster comparing to direct citation method, co-citation method, and co-word analysis (Boyack & Klavans, 2010; Ahlgren & Jarneving, 2008). LDA is a text mining method based on the Bayesian statistics. Latent research topics are identified by extracting terms with semantic similarity. As a relatively new method in scientometric, LDA and its extended methods like author-topic model have been used to analyse research similarity (Lu & Wolfram, 2012). Due to the difference in analysis objects, citations and abstracts of articles, we are interested in comparing the research topics that could be extracted from the two methods, and try to distinguish which method performs better. Experts' opinions are always used for evaluating the cluster quality; however in this case, discussions on IDR probably scatter in many research areas. Therefore, it would become the challenge for this study.

Results

First, we analyse the topics identified by each algorithm, and merge the categories with similar themes. Around 15 clusters discussing IDR have been identified from each method. They include discussing IDR from the aspect of history of science and philosophy of science, interdisciplinary medicine, interdisciplinary education, evaluating on interdisciplinary from the perspective of scientometrics and bibliometrics, knowledge management. We consider the remaining categories as general interdisciplinary research, which implies that they promote the use the IDR ideas or methods solving such problems as immigration and cultural propagation; sustainable development and environment, and ecological problem; general business management problems like design and service; juvenile delinquency; complex network problem etc.

We found that topics generated from each algorithm are basically unanimous, only with one exception that knowledge management are classified as a perspective of IDR by Louvain method, while LDA method does not identify this topic. That may due to LDA defines the term "knowledge" as a stop word automatically, which implies it considers "knowledge" as a common word and cannot provide any meaningful information for the investigation.

Second， although this finding reveals us the discussions on current interdisciplinary research, it does not shed light on more detailed information contained in each cluster. To explore characteristics of the cluster more deeply, we re-examine the frequent authors and frequent references of the cluster to provide a more precise and powerful interpretation. Due to restrict of specialty, we could not give detailed

analyse on each cluster. In this case, we take the topic about measurement and evaluation on IDR generated form LM algorithm as an example. From most frequent references, it seems that some publications are not relevant to IDR like Social network analysis, Introduction to modern information retrieval. We think this happens because knowledge related to network analysis and informetrics are applied to measure and evaluate the interdisciplinary research. Therefore, some publications about network analysis and information retrieval study were frequently cited by articles in this cluster.

Publications written by Leydesdorff L., Porter A.L. and Morillo F. are widely cited in this cluster. And researchers like Leydesdorff L., Porter A.L. and Rafols, I. are the top 3 active authors. That is in accordance with our understanding of this cluster.

Table 1. Most frequent references and authors

Most frequent references	Most frequent authors
Wasserman S, 1994, Social network analysis: Methods and applications	Leydesdorff, L
Salton G, 1983, Introduction to modern information retrieval	Porter, AL
Boyack KW, 2005, Mapping the backbone of science	Rafols, I
Ahlgren P, 2003, Requirement for a cocitation similarity measure, with special reference to Pearson's correlation coefficient.	Fegert, JM
Morillo F, 2003, Interdisciplinarity in science: A tentative typology of disciplines and research areas.	Ziegenhain, U

Third, to compare which method performs better, we dig deeper on the topic about measurement and evaluation on IDR. According to LM, 61 articles are included in this cluster. However, after reading abstracts of each article, we found around 20 articles are not about this topic. Then, we sorted the articles according to their probabilities of measurement and evaluation on IDR. Through analysing abstracts of the 61 top articles, we found that 55 of these articles are certainly associated with measurement and evaluation on IDR. More systematically comparisons will be conducted in future research.

Reference

Blondel, V.D., Guillaume, J.-L., Lambiotte, R. & Boyack, K.W. & Klavans, R. (2010). Co-Citation Analysis, Bibliographic Coupling, and Direct Citation: Which Citation Approach Represents the Research Front Most Accurately? *Journal of the American Society for Information Science and Technology*, 61(12), 2389-2404.

Committee on Facilitating Interdisciplinary Research and Committee on Science. (2005) *Engineering, and public policy, facilitating interdisciplinary research, the national academies.* Washington, DC: The national academies press.

Kessler M.M. (1963). Bibliographic coupling between scientific papers. *American Documentation*, 14(1), 10-25.

Lu K. & Wolfram D. (2012). Measuring Author Research Relatedness: a comparison of Word-Based, Topic-Based, and Author Cocitation Approaches. Journal of the American Society for Information Science and Technology, 63(10), 1973-1986.

THE STUDY AND ASSESSMENT OF RESEARCH PERFORMANCE AT THE MICRO LEVEL: THE AGE PHASE DYNAMICS APPROACH

Victor Rybachuk[1*] and Galena Quist[2]

[1] *vnyugvpr@yandex.ru; rybachuk.v.p@nas.gov.ua;* [*]*corresponding author*
The National Academy of Sciences of Ukraine, G.M. Dobrov Center for Science & Technology Potential and Science History Studies, 60 Blvd T. Shevchenko, 01032 Kyiv (Ukraine)

[2] *gvqr@live.com*
Ghent University, Salisburylaan 133, B-9820 Merelbeke, Ghent (Belgium)

Introduction

One of the central problems in micro level evaluation of scientific activity effectiveness is the relationship of productivity and quality of scientific publications to the age of the scientist. Despite the progress achieved in this field, the topic still remains actively researched and debated (see for example Bonaccorsi & Daraio, 2003; Costas, van Leeuwen & Bordos, 2010; Hörlesberger, Holste, Schiebel et all, 2011; Reijnhould, Costas, Noyons, Boerner & Sharnhorst, 2013). In the present study, investigations of the relationship between scientist's age and their publication activity are discussed in the context of Age Phase Dynamics (APhD) model of scientific performance (Pelz & Andrews, 1966; Malitsky, 1988; Rybachuk, 2005).

Methodology

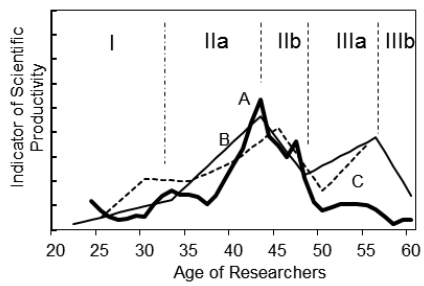
Pelz & Andrews (1966) found a bimodal dependence between productivity and age of the scientists. They divided the life cycle of scientific performance into individual phases and highlighted the main factors responsible for the wave-like age dynamics of the scientist's publication activity. Fox (1983) interpreted the discovered wave-like pattern to be a consequence of authors' considering not only the selected major works of the scientist, but a wider spectrum of publications, including articles, patents, presentations, as well as manuscripts. Later, Malitsky (1988) defined the basis of the model proposed by Pelz and Andrews, reformulated and further developed it to become a "principle of phase dynamic development of researcher's scientific activity". In our opinion, the age phase dynamics approach integrates the elements of econometric models and models of human capital, in terms of the sociology of science.

Method

The main criteria for selection of bibliographies for analysis were all-inclusive coverage of publications and the availability of their bibliometric data. Unpublished materials, electronic and media publications were not included. References in analyzed bibliographies (not shown) were confirmed in Scopus and Google Scholar scientometric databases.

Results and Discussion

This poster presents some preliminary results of age phase dynamics analysis (APhD-analysis) of the scientific bibliographies of 14 renowned scientists. The selected scientists conducted research in a different field from each other and during different historical periods. APhD-analysis of bibliographies of four representative scientists are shown in Figures 1 and 2.



Types of Knowledge Movement	Phases of Development of Scientific Activity				
	I	IIa	IIb	IIIa	IIIb
Accumulation	+	+	+		
Production		+	+	+	
Transmission			+	+	+

Figure 1. Age dynamics of publications of T.S. West (Kalyane & Munnolli, 1995) (A) in comparison with productivity obtained for groups of scientists by B. Malitsky (1988, p. 95, Fig. 11) (B) and by D. Pelz and F. Andrews (1973, Rus. Ed., p. 285, Fig. 57) (C). (The scientific productivity indicator axis is approximate).

Figure 1 shows the curve of a British chemist, Thomas S. West (1927-2010) as compared to those of other scientists published by Malitsky (1988) and Pelz & Andrews (1966). Figure 2 shows the curves of an American, 2001 Nobel Laureate in Physiology/Medicine, Leland H. Hartwell (born 1939); a French 1991 Nobel Prize laureate in Physics, Pierre-Gilles de Gennes (1932-2007); and a Ukrainian economist, Gennady Dobrov (1929-1989). The productivity curves in each phase of scientific performance life-cycle clearly indicate the most common type of scientific activity (movement of knowledge), and the nature of scientific and organizational functions of the scientists (performing, guiding, training, consulting) in the corresponding period of their career (see table in Fig. 1). Notable is that Phase 1 in the publication activity of many authors is usually unremarkable.

The indicated general pattern of APhD of scientific activity (Fig. 1) may be complicated by the influence of various non-systemic factors that reflect subjective circumstances and external conditions during the scientist's career. In particular, the nature of the research (theoretical or experimental), the field of science, the change of scientific focus, following the principle of "Publish or Perish", and so forth may alter the pattern.

Fig. 2 illustrates the differences in the types of APhD as related to the researcher's field of science: Physics (G), Molecular Biology and Genetics (H), and Economics and Sociology of Science (D).

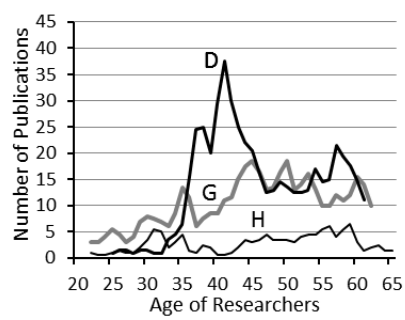


Figure 2. APhD of publications of G.M. Dobrov (curve D), L.H. Hartwell (H) and P.-G. de Gennes (G). Graphs are shown in the form of trends of linear filtration method with a period of 3 years.

We attempted to establish classification criteria for individual APhD-profiles and considered their further integration within the publication profiles of small research groups and laboratories as well as large groups of scientists (meso level), also incorporating the interrelationship with citation indices (data not shown). Encouraging in this respect, are the data recently published by Costas, van Leeuwen, & Bordons (2010). Age profiles of a total number of publications and citations per publication (Fig. 6 and Fig. 7 in referenced article) that they obtained at meso level are in good agreement with the typical individual APhD-profiles at micro level, presented in this work.

Further Research

This work was carried by G.M. Dobrov Center for S&T Potential and Science History Studies of Ukrainian NAS, as part of the BILAT-UKRAINA project within the European Commission FP7-INCO-2012-2.2 grant agreement 311839. As part of this project, we plan to evaluate the effect of international scientific cooperation, in particular that of Ukraine and the EU, on the APhD-profiles of scientific activity.

References

Bonaccorsi, A., & Daraio, C. (2003). Age effects in scientific productivity. *Scientometrics*, *58*(1), 49-90.

Costas, R., van Leeuwen, T. N., & Bordons, M. (2010). A bibliometric classificatory approach for the study and assessment of research performance at the individual level: The effects of age on productivity and impact. *Journal of the American Society for Information Science and Technology*, *61*(8), 1564-1581.

Fox, M. F. (1983). Publication productivity among scientists: A critical review. *Social Studies of Science*, *13*(2), 285-305.

Hörlesberger, M., Holste, D., Schiebel, E., Roche, I., Francois, C., Besagni, D., & Cuxac, P. Measuring the Preferences of the Scientific Orientation of Authors from their Profiles of Published References. In *European Network of Indicator Disigners* (7-11September, 2011). Rome, ENID.

Kalyane, V. L., & Munnolli, S. S. (1995). Scientometric portrait of T.S. West. *Scientometrics*, *33*(2), 233-256.

Malitsky, B.A. (1987). Phase dynamics of scientific activity and impact of scientist's efforts / In V.E. Tonkal & G.M. Dobrov Eds*., Scientific and technical potential: Structure, dynamics, efficiency* (pp. 88-101). Kiev: Naukova Dumka.

Pelz, D. C., & Andrews, F. M. (1966). *Scientists in organization: Productive climates for research and development*. J. Wiley.

Pelz, D., & Andrews, F. (1973). *Scientists in organizations: Optimal conditions for research and development* (471 pp.). *Ed. in Russian,* Moscow: Progress.

Reijnhoudt, L., Costas, R., Noyons, E., Boerner, K., & Scharnhorst, A. (2013). " Seed+ Expand": A validated methodology for creating high quality publication oeuvres of individual researchers. *arXiv preprint arXiv:1301.5177*.

Rybachuk, V.P., Grachev, O.A., Kukhtenko, T.A. & Videnina, N.G. (2005). Defining the general and specific bibliometric characteristics of research activities of scientists. *Nauka ta Naukoznavstvo (Science and Science of Science)*, 4 (Addendum), 105-112.

THE SUBJECT CATEGORIES NORMALIZED IMPACT FACTOR

Pablo Dorta-González[1] and María-Isabel Dorta-González[2]

[1] *pdorta@dmc.ulpgc.es*
Universidad de Las Palmas de Gran Canaria (Spain)

[2] *isadorta@ull.es*
Universidad de La Laguna (Spain)

Introduction

During decades, the journal impact factor (IF) has been an accepted indicator in ranking journals. However, there are increasing arguments against the fairness of using the IF as the sole ranking criteria (Dorta-González & Dorta-González, 2010, 2011, 2013). These indicators are not comparable among fields of science for two reasons: (i) each field has a different impact maturity time, and (ii) because of systematic differences in publication and citation behaviour across disciplines. Therefore, citation-based indicators need to be normalized for such differences.

There are statistical patterns which are field-specific and allow the normalization of the IF. Garfield proposes the term 'citation potential' for systematic differences based on the average number of references per paper. The fractionally counted IF corrects these differences in terms of the sources of the citations. Zitt & Small (2008) propose the *Audience Factor* using the mean of the fractionally counted citations to a journal. Similarly, Moed (2010) divides a modified IF by the median number of references in the Scopus database. He proposes the resulting ratio as the *Source Normalized Impact per Paper* (Leydesdorff & Opthof, 2010).

In addition to the average number of references, there exist some other sources of variance. In this work we decompose the aggregated impact factor into five main sources of variance and calculate them in all the categories of the JCR. Furthermore, a normalization process considering all journals in the indexing categories is proposed.

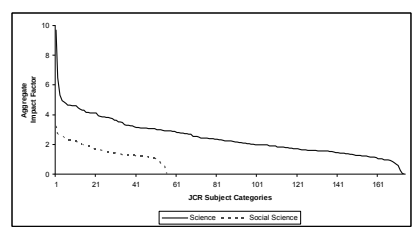


Figure 1. Aggregate Impact Factor of the JCR subject categories

Decomposing the Aggregate Impact Factor

Let F be the set of all journals in a specific field. Then, the *Aggregate Impact Factor* (AIF) is the ratio between the citations in year t to citable items in any journal of field F in years t-1 and t-2, and the number of citable items published in those years.

It is possible to decompose the AIF into five main variables: field growth rate, average number of references, ratio of

references to JCR items, ratio of JCR references to the target window, and proportion of cited to citing items in the target window. We prove that these variables are normally distributed and different across fields.

Categories Normalized Impact Factor (CNIF)

Let $F_t^1, F_t^2, ..., F_t^n$ be the subject categories where journal *i* is indexed in year *t*.

Denoting by $\bigcup F_t^j = F_t^1 \bigcup F_t^2 \cdots \bigcup F_t^n$, then

$$AIF_t^{\bigcup F_t^j} = \sum_{i \in \bigcup F_t^j} NCited_t^i \Big/ \sum_{i \in \bigcup F_t^j} A_{t-1}^i + A_{t-2}^i .$$

In a similar way,

$$AIF_t^{JCR} = \sum_{i \in JCR} NCited_t^i \Big/ \sum_{i \in JCR} A_{t-1}^i + A_{t-2}^i .$$

We define the *Categories Normalized Impact Factor* of journal *i* in year *t* as:

$$CNIF_t^i = IF_t^i \cdot \left(AIF_t^{JCR} \Big/ AIF_t^{\bigcup F_t^j} \right).$$

Therefore, this indicator has an intuitive interpretation, similar to the IF.

Materials and Methods

The bibliometric data was obtained from the online version of the Journal Citation Reports (JCR) during the first week of October 2011. The 2010 Science edition contains 8,073 journals classified into 174 subject categories, and the Social Science edition contains 2,731 journals classified into 56 subject categories.

In the comparative analysis between the IF and the CNIF, and the estimation of the gap between rankings across categories, the five selected categories are: Business; Business, Finance; Economics; Management; Operations Research & Management Science. These categories contain a total of 590 different journals (490 in just one category, 98 in two, and 2 in three).

Results and discussion

The AIF in Science is around 58% higher than in Social Science (see Figure 1). This is due to the fact that although on average there are over 30% more references in articles of Social Science, an important part of them are non-JCR items. In Social Science around 40% of the references are books and journals that are not indexed in the JCR, while in Science these are around 20% of the references.

The categories with the highest AIF in Science are related to biomedicine. The lowest values are in engineering and mathematics. In Social Science, the categories with the highest values are related to psychology and some specialties of economics, such as health policy and management. The lowest factors are in categories related to history.

A journal from a category of Social Science has on average 30% more references and around 20% more references to non-JCR items than a journal from a Science category. The longest reference lists are produced in history and the shortest in engineering and mathematics. The highest proportions to non-JCR items are in physics, biology, and chemistry, whereas the lowest are in engineering and computer science.

One of each five JCR references is on average in the target window. Curiously, some of the categories with the lowest proportion of references to JCR items have the highest proportion of citations in the target window. This happens, for example, in engineering and history. In some areas, such as mathematics, just one in eight JCR references is from the previous two years, in comparison to history where they are one in three.

With respect to the Cluster Analysis of the JCR categories, C1 and C7 include, in general, those life sciences with an

important social component, as well as those social sciences which use mathematical methods in a higher degree (health, psychology, economics, and business, for example).

Clusters C2 and C4 contain those social sciences which use mathematical methods in a lower degree (education, sociology, linguistic, and law, for example). Finally, clusters C5 and C6 include formal, physical, technological, and life sciences (mathematics, physics, chemistry, engineering, and biomedicine, for example).

The variance in the AIF in Science can be explained to a great degree by three major components (the ratio of references to JCR items, the ratio of JCR references to the target window, and the field growth). In Social Science, the variance can be explained to a great degree by only two major components (the ratio of JCR references to the target window and the proportion of cited to citing items in the target window). The principal components are different depending on the edition of the JCR. This is motivated because in Social Science there are many different disciplines in relation to the habits of publication and citation (e.g. economics and psychology versus history).

Finally, there are important differences between the CNIF and the IF for most of the journals analyzed. In the case of the CNIF, the maximum gap is reduced in more than half of the journals with respect to the IF. The average gap is also reduced by around a 32%.

Conclusions

A decomposing of the field aggregate impact factor into five normally distributed variables shows that, for the JCR subject categories, the variables that to a greater degree explain the variance in the impact factor of a field do not include the average number of references. However, this is the factor that has most frequently been used in the literature to justify the differences between fields of science, as well as the most employed in the source-normalization (Moed, 2010; Zitt & Small, 2008). Therefore, it is necessary to consider some other sources of variance in the normalization process.

References

Dorta-González, P., & Dorta-González, M. I. (2010). Indicador bibliométrico basado en el índice h. *Revista Española de Documentación Científica*, 33(2), 225–245.

Dorta-González, P., & Dorta-González, M. I. (2011). Central indexes to the citation distribution: A complement to the h-index. *Scientometrics*, 88(3), 729–745.

Dorta-González, P., & Dorta-González, M. I. (2013). Comparing journals from different fields of science and social science through a JCR subject categories normalized impact factor. *Scientometrics* (in press). DOI 10.1007/s11192-012-0929-9

Leydesdorff, L., & Opthof, T. (2010). Scopus's source normalized impact per paper (SNIP) versus a journal impact factor based on fractional counting of citations. *Journal of the American Society for Information Science and Technology*, 61(11), 2365–2369.

Moed, H. F. (2010). Measuring contextual citation impact of scientific journals. *Journal of Informetrics*, 4(3), 265–277.

Zitt, M., & Small, H. (2008). Modifying the journal impact factor by fractional citation weighting: The audience factor. *Journal of the American Society for Information Science and Technology*, 59(11), 1856–1860.

SUCCESS DETERMINANTS OF FULL-TIME RESEARCHERS AT HOSPITALS. A PERCEPTIONS-BASED STUDY

Jesús Rey-Rocha[1], Belén Garzón-García[1], Irene López-Navarro[1] and M. Teresa Antonio-García[2]

[1]*jesus.rey@csic.es, irene.lopez@cchs.csic.es*
Research Group on Scientific Evaluation and Transfer. Spanish Council for Scientific Research (CSIC). C/ Albasanz, 26-28. Madrid (Spain)

[2] *mantonio@.bio.ucm.es*
Research Group on Scientific Evaluation and Transfer (CSIC) & Universidad Complutense de Madrid. Faculty of Biological Sciences. Ciudad Universitaria. Madrid (Spain)

Introduction

Different factors affecting performance and productivity of researchers have been described in the literature. Namely individual factors (Bonaccorsi & Daraio, 2003; Fox, 2005; Leahey , 2006; van Arensbergen et al., 2012,), contextual and organizational factors (Smeby & Try, 2005; Seashore et al., 2007), and psychological factors (Rey-Rocha et al., 2007; Torrisi, 2013).

Most studies have been carried in an academic environment, mainly in laboratories. But these factors may affect researchers activity in a different way within the essentially clinical hospital environment. In this work we investigate the extent to which different individual and institutional characteristics can influence performance and productivity of researchers within the hospital setting.

The Miguel Servet (MS) Research Contract Programme is one of the most important strategic actions being undertaken by Spanish Administration in order to enhance the research activity at public hospitals. The Programme is aimed at incorporating researchers with excellent training within the National Health System (NHS) in order to improve its research capacity and to promote the creation of stable research groups within the NHS.

Methodology

Population, sample and research instruments

The universe to be studied consisted of the 367 researchers funded by first eigth calls (1998-2005) of the MS Programme, whose contracts ended between 2005 and 2012.

We used a web-based survey to obtain data from the population of MS researchers (72.2% response rate). Data on research activity and productivity were obtained from the activity reports submitted by researchers.

The present work is based on data from the 174 researchers who finished its six-year contract and who answered the survey.

Variables

After the six-year contract, MS researchers' activity and results are

evaluated anew for those who wish to apply for a further five-year contract through the Researcher Stabilization Programme. To be evaluated positively, researchers must demonstrate a certain productivity in high impact journals together with leadership (i.e. leading of funded research projects and first authorship of articles).

Thus, in this work research performance of researchers has been assessed through the following indicators:

- art_N: number of articles in ISI journals.
- art_Q1, %art_Q1: number and percentage of articles in first-quartile ISI journals.
- art_FL, %art_FL: number and percentage of ISI articles as a first or last author.
- proj_N: number of funded projects.
- proj_PR, %proj_PR: number and percentage of projects as principal researcher.

Researchers were asked about different aspect of their research activity and their perceptions, judgements, thoughts and feelings about this activity and its organizational context. In this paper we investigate the effect of the following factors:

a) Satisfaction with… (in a 1 to 5 scale):
   - Scientific quality of the host group.
   - Scientific quality of the host centre.
   - Research autonomy.
   - Decision-making capacity.
   - Leadership.
   - The conditions of the facilities and space available.
   - Job stability expectations.
b) Satisfaction with the resources at their disposal (1 to 5):
   - Human resources: technical and support staff and researchers in training.
   - Material resources: infrastructures, equipment and research materials.
   - Support units.
   - Economic resources.
c) Creation of new research groups (Yes, my incorporation has led to the creation of a new research group I lead / No, I stayed in a already existing group).
d) Self-assessment of their contribution to the relationship between clinical and basic researchers (1 to 5).
e) Type of research performed (basic, clinical, both).

Data analysis

In order to determine whether the means for paired samples were systematically different, we applied the Student's t-test, adjusted using the Bonferroni correction.

Results

Productivity and the capacity to obtain research projects are related with researchers' satisfaction with the human resources in their groups. Thus, art_N increases by 57% in satisfied versus unsatisfied researchers. The capacity to publish in top journals is also influenced by this satisfaction: art_Q1 increased by 65% (Figure 1). Likewise, satisfied researchers participated in 44% more projects than those unsatisfied, but did not obtain a significant higher number of projects as principal researcher.

As expected, leadership of a research group increases proj_PR and %proj_PR (+ 61% and +29% respectively).

Productivity in ISI journals is also related with the kind of research performed. Researchers doing clinical research published more articles (65% more than those doing basic research and 21% more than basic+clinical researchers), more art_Q1(+ 70% than basic) and obtained a higher proj_N and

proj_PR (+69% and +98% respectively) (Figure 2).

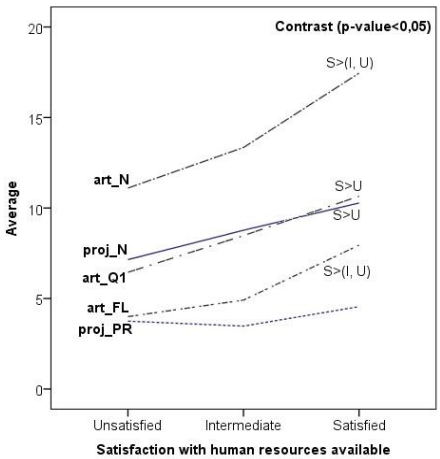


Figure 1

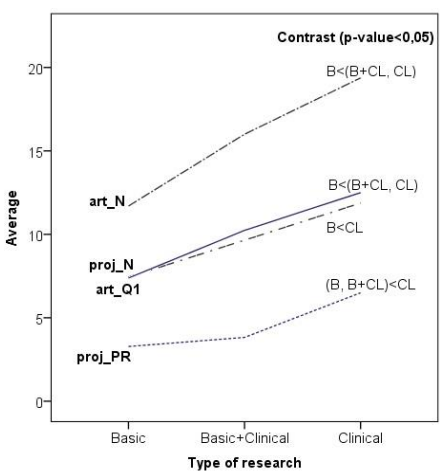


Figure 2

Acknowledgments

This work was supported by the Spanish Ministry of Health, within the framework of the Spanish RDI Plan (grant numbers PI06/0983 and PI10/00462).

References

Bonaccorsi, A. & Daraio, C. (2003) Age effects in scientific productivity. The case of the Italian National Research Council (CNR). *Scientometrics*, 58(1), 49-90.

Fox, M.F. (2005). Gender, family characteristics and publication productivity among scientists. *Social Studies of Science*, 35(1), 131-150.

Leahey, E. (2006). Gender differences in productivity: research specialization as a missing link. *Gender and Society*, 20 (6), 754-780

Rey-Rocha, J., Garzón-García, B. & Martín-Sempere, M.J. (2007). Exploring social integration as a determinant of research activity, performance and prestige of scientists. Empirical evidence in the Biology and Biomedicine field. *Scientometrics*, 72, 59-80.

Seashore, K., Holdsworth, J.M., Anderson, M.S. & Campbell E.G. (2007). Becoming a Scientist: The effects of work-group size and organizational climate. *The Journal of Higher Education*, 70(3), 311-336.

Smeby, J.C., Try, S. (2005). Departmental contexts and faculty research activity in Norway. *Research in Higher Education*, 46 (6), 593-619.

Torrisi, B. (2013) Academic productivity correlated with well-being at work. *Scientometrics*, 94, 801-815.

Van Arensbergen, P, van der Weijden, I. & van den Besselaar, P. (2012). Gender differences in scientific productivity: a persisting phenomenon? *Scientometrics* ,93, 857-868.

SURFING THE SEMANTIC WEB

Sojung Yang , Hyosook Jung, and Seongbin Park*

{cherry089, est0718, hyperspace}@korea.ac.kr
** corresponding author*
Korea University, Seoul, Korea

Introduction

In this paper, we address the structural issues that can affect how a user can surf the Semantic Web (Antoniou & van Harmelen, 2004). Like the Web, the structure of the Semantic Web is a hypertext and users still need to follow hyperlinks to access the information (Goble, Bechhofer, Carr, De Roure, and Hall, 2001). However, it is not always easy to find the desired information in a large scale hypertext environment by traversing hyperlinks (Zhou, Leung, and Winoto, 2007). The motivation of our research is that certain structures such as a matroid (Cormen, Leiserson, Rivest, and Stein, 2003) and a small world network (Watts, 1999) have been proven useful in constructing a navigable Web site, where navigability refers to how easy a user can find desired information as the user freely moves around at a Web site (Min, Chun, Jang, Jung, and Park, 2011).

Two structures

A matroid is a pair (S, I), where S is a finite set and I is a set of subsets of S which satisfies the following two conditions: (1) If A is an element of I and B is a subset of A, then B is also an element of I. (2) If A and B are elements of I and A contains more elements than B, then there exists an element x in A - B such that $\{x\} \cup B$ is an element of I (Cormen et al., 2003). One motivation of considering a matroid structure lies in the fact that a computational problem that exhibits a matroid structure can be solved using a greedy algorithm (Cormen et al., 2003) and if a network can be constructed as a matroid, greedy browsing (i.e., each time a user visits a web page, the user visits what looks most relevant to the desired information using the local information only) may lead to a shortest path from the current location to a destination. To demonstrate this idea, let's assume that we build a web site about history of painting using 16 web pages, where each page contains the information about 17C, 18C, 19C, 20C, Baroque, Rococo, Neoclassicism, Realism, Impressionism, Cubism, Rembrandt, Boucher, David, Millet, Van Gogh, and Picasso, respectively. We can build a web site whose structure is a slight modification of a matroid structure and reflects ontological constraints, where an ontology represents vocabularies and their relationships in a domain of interests (Staab & Studer, 2009). We made a modification because if we consider all the conditions in the definition of a matroid, the number of hyperlinks becomes too big.

Let S be the set of 16 web pages. Then we include subsets of S in I as follows:(1) Include a standard of classification in each element of I. With this criterion, I = {{17C}, {18C}, {19C}, {20C}}. (2) Associate one painting movement with the corresponding standard of classification. With this criterion, I contains elements

such as {17C, Baroque}, {18C, Rococo}, etc. (3) Associate one art movement with one painter. With these criteria, I contains the following elements: {17C, Baroque, Rembrandt}, {18C, Rococo, Boucher}, {19C, Neoclassicism, Fuseli, Realism, Millet}, {20C, Impressionism, Van Gogh, Cubism, Picasso}. Members in an element of I are connected one another by hyperlinks. One interpretation of the resulting structure is that it is a collection of subsets of S, where each subset contains some representative element (in our example, 17C, 18C, etc) and other representative element (Baroque, Rococo, etc) etc. Finally, we add additional hyperlinks that come from ontological constraints. For example, a link between "Rembrandt" and "Van Gogh" can be created if "Rembrandt" and "Van Gogh" have the same value of property "is_From" that is "The Netherlands" in an ontology. Similarly, Boucher and Millet can be linked if their value of property "is_From" is the same ("France"). "Van Gogh" and "Millet" can be linked as well. With these additional constraints, two elements which belong to different subsets of S can be connected by hyperlinks. Figure 1 shows the resulting structure.

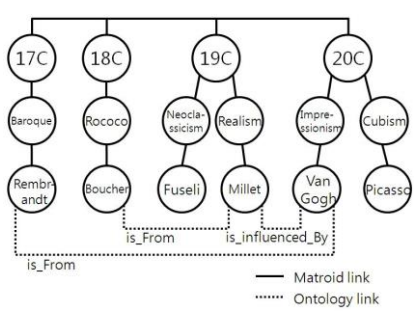


Figure 1. The structure of a web site with matroidal and ontological constraints.

A small world network has a relatively small average path length compared to the number of nodes in the network (Watts, 1999). One motivation of considering a small world network is that if a network forms a small world, efficient routing is possible (Kleinberg, 2000). Routing in a network is very similar to traversing hyperlinks that connect web pages in the sense that both of them involve following a sequence of links with only local information available in each step. One way to construct a small world network is to arrange the nodes in a lattice and connect two arbitrary nodes with the probability that is proportional to one over the square of the lattice distance between the nodes (Costa & Barros, 2006). Assuming that we have the same set of pages in the previous example, we first connect the standards of classification according to the order of periods. Each period is connected with a painting movement and each painting movement is connected with a painter. Then, two nodes are connected based on the following criteria. If the relationship between two nodes is defined in an ontology, then connect them. Otherwise, connect them with the probability which is inversely proportional to its lattice distance. Figure 2 shows a resulting structure. The probability that there is a link between "Van Gogh" and "Rembrandt" is 1/16 because their lattice distance is 4. However, a direct link between the nodes can be created if the relationship is defined in an ontology.

Semantic Web browser

One way to view the Semantic Web is that it consists of logical theories (Sipser, 2006) where sentences correspond to RDF triples (Klyne & Carroll, 2004) and hyperlinks exist among the set of sentences. A Semantic

Web browser should be able to parse and derive new facts while a user is surfing the Semantic Web. Derived facts correspond to logical consequences of the sentences (i.e., RDF triples) that the user has been visiting. In addition, a Semantic Web browser should be able to infer implicit relations such as transitivity relation. It is also desirable that a Semantic Web browser provides adaptive links based on navigation history. For example, in figure 3, if a user visits a node "The Netherlands" from a node "Rembrandt" and then visits a node “The Netherlands" from a node "Van Gogh", then it can help if the browser provides hyperlinks (dotted links in the figure) to a node "Van Eyck" and a node "Escher".

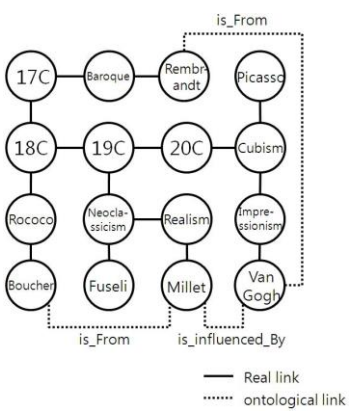


Figure 2. The ontological small world.

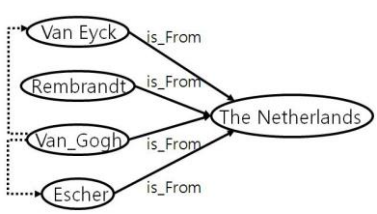


Figure 3. The structure of web pages.

Conclusion and ongoing works

In this paper, we addressed the issue of the navigability of the Semantic Web. Currently, we are working on ways by which users can exploit structural properties to surf the Semantic Web in a greedy way. We plan to measure the navigability of the structures that reflect (1) matroid and ontological constraints (2) small world property and ontological constraints.

References

Antoniou, G. & van Harmelen, F. (2004). A Semantic Web Primer, The MIT Press.

Cormen, T. H., Leiserson, C. E., Rivest, R. L. & Stein, C. (2003). Introduction to Algorithms, MIT Press.

Costa , A. R. & Barros, J. (2006). Network information flow in navigable small-world networks, Proceedings of the 4th international symposium on modeling and optimization in mobile, ad hoc and wireless networks.

Goble, C., Bechhofer, S., Carr, L., De Roure, D. & Hall, W. (2001). Conceptual Open Hypermedia = The Semantic Web?, *Semantic Web Workshop*.

Kleinberg, J. M. (2000). Navigation in a small world, *Nature,* Vol 406.

Klyne, G. & Carroll, J. J. (Ed.) (2004). Resource Description Framework (RDF): Concepts and Abstract Syntax. Retrieved January 26, 2013 from: http://www.w3.org/TR/rdf-concepts/# section-triples

Min, K., Chun, S., Jang, G., Jung, H. & Park, S. (2011). The structure of a web site and navigability, *The Journal of Korean Association of Computer Education,* Vol 14, No 3.

Sipser, M. (2006). Introduction to the theory of computation, 2nd edition, Thomson Course Technology.

Staab, S. & Studer, R. (2009). Handbook on Ontologies, Springer Publishing Company.
Watts, D. J. (1999). Networks, Dynamics, and the Small-World Phenomenon, *The American Journal of Sociology*, Vol 105, No 2
Zhou, Y., Leung, H. & Winoto, P. (2007). MNav:
A Markov Model-Based Web Site Navigability
Measure, Journal of IEEE Transactions on
Software Engineering, volume 33, issue 12.

TEMPORAL EVOLUTION, STRUCTURAL FEATURES AND IMPACT OF STANDARD ARTICLES AND PROCEEDINGS PAPERS. A CASE STUDY IN BLENDED LEARNING.

Barrios, Maite[1], González-Teruel, Aurora[2], Cosculluela, Antonio[1], Fornieles, Albert[3], Ortega, Lidia[1], Turbany, Jaume[1]

[1] *mbarrios@ub.edu, acosculluela@ub.edu, lortegca8@gmail.com, jturbany@ub.edu*
Dept. of Methodology of Behavioral Sciences. University of Barcelona, Barcelona (Spain)

[2] *Aurora.Gonzalez@uv.es*
Dept. of Science of history and Documentation. University of Valencia. Valencia (Spain)

[3] *albert.fornieles@uab.cat*
Dept. of Psychobiology and Methodology of Health Sciences. Autonomus University of Barcelona. Barcelona (Spain)

Introduction

Inevitably, teaching and learning in higher education have been transformed by the arrival of new computer-based technologies (Newman, Couturier & Scurry, 2004) and by the opportunities they provide for addressing the needs of society in the twenty-first century. Blended learning emerged in the late 1990s as a term to refer a new approach in education that mixes traditional face-to-face instruction with online learning (Garrison & Vaughan, 2008). Since then, the universities that have incorporated blended learning into their courses have expanded enormously (Arabasz & Baker, 2003).

This paper studies the development of scientific production since the introduction of blended learning by focusing on proceedings papers (PP) and standard articles (SA). In the study of scientific communication, the role of PP may differ depending on the field (Drott, 1995, Lisée, Larivière & Archambault, 2008). In some disciplines (e.g., engineering) they may be substitutes for journal papers, while in others they represent preliminary material that will later be developed into a more rigorous manuscript for publication in a journal (Drott, 1995, Glänzel, Schlemmer & Schubert, 2006, Lisée, et al., 2008). However, the presence and the relevance of PP in new emerging fields have not been explored in depth. We feel that the study of blended learning since its origins can provide a picture of the relevance of PP in the evolution and consolidation of a new educational approach. In this study we analyse the publication rates of SA and PP in a new and emerging field in order to compare and contrast their temporal evolution, structural features (number of co-authors, number of references and pages) and impact (assessed by the number of citations).

Method

Data collection

The documents included in the present study were identified using the Thomson Reuters Web of Science (WoS) database. In order to retrieve the relevant scientific literature, the search was performed in the topic field (which runs the search for titles, keywords and abstract) and using the truncated form of different synonyms used to refer to blended learning methodology (i.e., blended e-learning, b-learning, m-learning, hybrid, flip, mixed-mode, web-enhanced, technology mediated), synonyms referring to the learning process (i.e., learning, instruction, course, teaching, education) and other related words (i.e., face to face, distance, online). A total of 6,044 papers were initially retrieved. Titles and abstracts were checked manually to ensure that the paper actually implemented or reported a blended learning experience in higher education.

In order to study a paper's impact, the number of citations from its year of publication until the date of downloading was also obtained from the WoS database.

Data analysis

Descriptive statistics were used to study the temporal evolution of PP and SA. In order to study whether the growth in scientific output over time fitted Price's law, linear, exponential, logistic and cubic regression models were performed. The t-test was applied to determine whether SA and PP presented differences regarding structural features. Analysis of covariance (ANCOVA) was used to study differences in the number of citations corresponding to the two types of document while controlling for the effects of extraneous variables. As suggested by Bornmann, Mutz, Neuhaus, & Daniel (2008), the number of authors, pages per document and references were analyzed as covariates.

Results

Between 1998 and December 2012, a total of 1,260 documents dealing with blended learning in higher education were found in the WoS database. 55.3% (n= 697) of the documents were classified as PP and 44.7% (n = 563) as SA. Table 1 shows the proportion of variance explained for SA and PP.

Table 1. Regression fit of SA and PP

	SA	PP
R^2 linear	0.918	0.687
R^2 exponential	0.946	0.788
R^2 logistic	0.930	0.667
R^2 cubic	0.975	0.917

Figure 1 shows a clear upward trend in the publication rate during the 1990s and 2000s in both types of paper. In recent years, however, the publication rate of PP shows a clear downward trend.

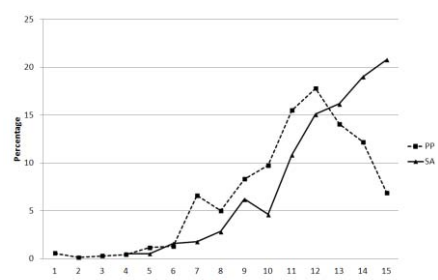


Figure 1. Temporal evolution in the number of publications of SA and PP

Significant differences were found between SA and PP in all variables measuring structural features. (Table 2).

ANCOVA showed that SA received a significantly higher number of citations than PP (SA (mean (SD): 4.50 (12.77),

PP: 0.41 (2.38); F(4, 1255) = 21.486, p < .001).

Table 2. Structural features of SA and PP

	SA	PP	p
	Mean (SD)	Mean (SD)	
Co-authors	2.98 (1.87)	2.58 (1.64)	<.001
References	32.90 (18.44)	14.07 (9.22)	<.001
Pages	11.72 (5.45)	7.21 (3.26)	<.001

Conclusions

The data show the importance of conferences and meetings in the emergence and consolidation of a new topic, in this case blended learning. In the initial period, PP were the most frequent type of document. As time passed, however, they were overtaken by SA. As for the differences in structural features, PP had fewer pages and references, presumably because these are limited by the conference rules or journal guidelines. The fact that the number of co-authors was higher in SA reflected a greater degree of collaboration than in PP. Judging from the number of citations received, SA were considered more relevant than PP. These results corroborate those of previous studies (Goodrum, MacCain, Lawrence & Giles 2001; Lisée, et al., 2008) and suggest that scientists regard SA as more elaborate and more mature than PP.

Acknowledgments

This study was supported by the research program in university teaching REDICE-12 (Redice12 1740-01) of the University of Barcelona.

References

Arabasz, P., & Baker, M. B. (2003). Evolving campus support models for e-learning courses. *Educause Center for Applied Research Bulletin.* http://www.educause.edu/ir/library/pdf/ecarso/ers/ERS0303/EKF0303.pdf.

Bornmann, L., Mutz, R., Neuhaus, C., & Daniel, H. D. (2008). Citation counts for research evaluation: Standards of good practice for analyzing bibliometric data and presenting and interpreting results. Ethics in Science and Environmental Politics, 8, 93–102.

Drott, C.M. (1995). Reexamining the role of conference papers in scholarly communication. *Journal of theAmerican Society for Information Science, 46*(4), 299–305.

Garrison, D. & Vaughan, N. (2008). Blended learning in higher education: Framework, principles, and guidelines. San Francisco, CA: John Wiley & Sons.

Glänzel, W., Schlemmer, B., Schubert, A., & Thijs, B. (2006). Proceedings literature as additional data source for bibliometric analysis. *Scientometrics, 68*(3), 457–473.

Goodrum, A. A., MacCain, K. W., Lawrence, S., & Giles, C. L. (2001). Scholarly publishing in the Internet age: A citation analysis of computer science literature. *Information Processing and Management, 37*(5), 661–675.

Lisée, C., Larivière, V., & Archambault, E. (2008) Conference Proceedings as a Source of Scientific Information: A Bibliometric Analysis. *Journal of the American Society for Information Science and Technology, 59*(11), 1776–1784.

Newman, F., Couturier, L., & Scurry, J. (2004). *The future of higher education: Rhetoric, reality, and the risks of the marker* . San Francisco: Jossey-Bass.

TESTING COMPOSITE INDICATORS FOR THE SCIMAGO INSTITUTIONS RANKING

Isidro F. Aguillo[1]

[1] *isidro.aguillo@csic.es*
Cybermetrics Lab. IPP-CSIC, Albasanz, 26-28, Madrid 28037 (Spain)

Introduction

The Scimago Institutions Rankings publish every year a global ranking of organizations all over the world performing scientific research. The last edition is the SIR World Report 2012 (http://www.scimagoir.com/pdf/sir_2012_world_report.pdf) that is one of the most prestigious rankings based on bibliometric-data.

Although many of the global rankings of universities use in a way or another bibliometric data, there are only a few of them that similarly to the Scimago's one are based solely on this information. These are the Leiden Ranking (http://www.leidenranking.com/) produced by CWTS in the Leiden University, the National University of Taiwan (NTU) Ranking (http://nturanking.lis.ntu.edu.tw/Default.aspx), that was formerly published by HEEACT and the University Ranking by Academic Performance (URAP) edited by the Turkish Middle East Technical University (http://www.urapcenter.org/).

Contrary to the Scimago Ranking that compiles information from Scopus/Elsevier database, they used instead the WoS/Thomson Reuters citation databases as a source.

NTU and URAP ranking only provides a unique classification based on a composite indicator that combines different bibliometric variables. Leiden ranking allows the end user to choose among different indicators that can be combined to produce 8 different rankings without supporting any of them explicitly. Scimago Ranking declares that is not a league table and in fact the ranking parameter (number of publications) is only for arrangement purposes and it is not a true ranking proposal. Even more they openly invite to use the report to rank institutions or to build a league table under your own responsibility.

The purpose of this contribution is to accept that invitation to explore and test several composite indicators alternatives built from the Scimago ranking published information.

The test will be performed against other ranking proposals so comparative analysis can be useful not only from an academic point of view but also in other applications of the rankings results.

Methodology

One important unique characteristic of the Scimago ranking is that includes not only universities but also a high number of other research - focused organizations (hospitals, research centers, private companies, International organizations, etc…). In order to make feasible the comparative analysis only the universities were selected.

The public report consists of seven indicators, namely Output (number of documents); International collaboration (output ratio with foreign co-authors); Normalized impact (normalized ratio

between institution's average scientific impact and the world average); High quality publications (ratio of publications in the first quartile journals); Specialization Index (thematic concentration / dispersion measured by Gini index); Excellence Rate (papers belonging to the 10% most cited ones in each discipline) and Scientific Leadership (papers with corresponding author belonging to the institution).

A first proposed indicator is not really a composite one as it is possible to rank the institutions multiplying their output by their normalized impact (ONI). This first indicator was used for building the list of top 500 universities that is the population study used for the analysis.

For the true composite ones we excluded international collaboration and specialization Index and considered (a) Output, Leadership and Q1 publications as activity related variables while Excellence Rate is an impact measurement, then (b) the ratio between activity and impact indicator should be 1:1 meaning their weighting parameters will be 50% each and finally (c) prior to combination the raw data will be log-normalized (ln(value+1)/ln(max+1)).

Table 1. Number of universities by country represented in the top 500 of the 2012 editions (2011 for Leiden) of the rankings.

Country	SCIM	LEID	URAP	NTU
us	139	127	129	154
de	39	39	40	46
uk	39	36	35	35
cn	36	26	34	23
it	26	25	25	27
ca	22	21	20	22
fr	28	20	15	20
jp	15	24	18	23
au	14	14	16	13
kr	11	18	16	12
es	14	16	14	13
nl	12	12	12	12
se	10	10	10	11
tw	8	9	10	6
ch	7	7	7	8
br	7	8	7	7
be	7	7	7	7
fi	6	6	6	6
il	5	6	6	5
at	6	5	5	6
pt	5	6	7	3
hk	5	5	5	5
gr	4	6	4	4
dk	5	4	4	4
za	3	4	5	3
no	3	3	4	4
ie	3	3	3	3
tr	1	6	5	
nz	2	3	4	2
pl	2	3	3	2
in	3	4	1	2
sg	2	2	2	2
cl	1	2	2	2
ir	3		4	
ru	1	2	2	1
ar	1	2	1	1
th		2	2	1
mx	1	1	1	1
cz	1	1	1	1
hr	1	1	1	1
hu		2		1
rs	1	1	1	
my			2	
si	1	1	1	1
sa			1	
eg			1	
ee			1	

Results

The population consists of universities of 41 different countries (considering Hong Kong as a separate unit), that it is close to the 40 countries from NTU or the 42 of the Leiden Ranking. Only URAP is slightly (46) more diverse. Table 1 summarizes the geographical distribution of the Top 500 universities in those rankings.

We obtained 4 indicators from Scimago data (including three composite ones) plus the global rank figures for URAP and NTU rankings. The 2012 Leiden ranking will be published shortly and

these figures will be incorporated in the final version.

The comparison between individual institutions were restricted in this short version to 16 different institutions representative from different regions as shown in Table 2

Table 2. Ranks of selected universities according to different indicators and sources: A: OUT*NC (Scimago); B: OUT+EXC (Scimago); C: LEAD+EXC (Scimago); D: Q1+EXC (Scimago); E: NTU; F: URAP

Domain	A	B	C	D	E	F
harvard.edu	1	1	1	1	1	1
stanford.edu	2	6	6	7	3	4
jhu.edu	3	5	5	2	2	3
utoronto.ca	6	2	3	5	7	2
ox.ac.uk	12	11	11	10	9	7
cam.ac.uk	14	13	13	13	15	11
u-tokyo.ac.jp	19	8	8	15	17	10
upmc.fr	26	20	31	22	48	41
ethz.ch	36	40	42	38	49	48
unimelb.edu.au	37	39	45	37	35	32
uu.nl	38	42	44	35	39	44
tsinghua.edu.cn	44	27	16	54	94	72
usp.br	53	35	33	55	53	28
tau.ac.il	129	109	117	106	116	93
uct.ac.za	278	287	286	276	280	237

Discussion

There are coverage differences between WoS and Scopus databases that can explain for example the better performance of the Tsinghua or the Paris VI universities (see Table 1 also). Other intra-Scimago discrepancies are mostly due to the factors considered when building the composite indicators. Some of them are very important as in the case of the Chinese university that is highly productive but with a comparative smaller impact. In any case all the results are wildly distinct from those offered in the current output-only list.

Recommendations

The results show that the several combinations can be useful for different purposes, providing rankings according to the end-users self-defined priorities. We strongly suggests that not only raw data can be provided as currently Scimago Ranking do, but also a series of candidate combinations are published at the same time. The criteria for combination proposed here are suggestions, but informed by previous expertise. Using an “a-priori” model with specific weightings and a normalization scheme that reduces outliers intends to avoid leaving open alternatives that could be meaningless.

A TEXT MINING APPROACH EXPLORING ACKNOWLEDGEMENTS OF PAPERS

Adrián A. Díaz-Faes and María Bordons

adrian.arias@cchs.csic.es, maria.bordons@cchs.csci.es
Instituto de Estudios Documentales sobre Ciencia y Tecnología (IEDCYT), Spanish National Research Council (CSIC), Albasanz 26-28, Madrid 28037 (Spain)

Introduction

Acknowledgements are an important aspect of scholarly communication, since they are used to recognize some special contributions to research that are not sufficient to qualify for authorship (Kassirer & Angell, 1991). As stated by Cronin, acknowledgement is a voluntary act that appears following an implicit code of professional conduct (Cronin, 1995). It has become a constitutive feature of the academic journal article in the 20th century, as well as a potentially rich source of insight into sub-authorship collaboration in science.

Until recently it was very difficult to carry out studies about acknowledgements (see for instance, Giles & Council, 2004) because this information was not available in bibliographic databases. However, the Web of Science (WoS) has been including funding acknowledgement data since August 2008. This database includes three fields of information on 'Funding Acknowledgement" (FA): 'Funding Agency' (FO) contains the names of the agencies that support the research, 'Grant Number' (FG) provides the numbers which identify the projects –if any- and 'Funding text' (FT) contains the full text included by the authors in the acknowledgement section of the paper. This recent development of the WoS database opens up new possibilities for a wide range of studies (Rigby, 2011), although the unstructured nature of the field complicates the analyses.

Acknowledgements can be studied with different purposes in science studies which range from the analysis of the interaction among scientists to their use in research evaluation and funding policy. Differences by disciplines in the frequency of acknowledgements and in the type of support acknowledged have been described (Costas and Van Leeuwen, 2012). In this paper we present a novel approach to explore acknowledgement patterns by disciplines and as a resource of sub-authorship information combining text mining and multidimensional data display techniques.

Data and method

Scientific papers published in four disciplines (Cardiac & Cardiovascular System, Economics, Evolutionary Biology, and Statistics & Probability) by Spain in 2010 were downloaded from WoS. The disciplines differ in their broad area ascription, theoretical/experimental orientation and basic/applied nature of research. These features are taken into account under the assumption that they might have an influence on the type of information to be included in the FA field. Textual analysis is developed on the information

provided by FT field. Orthographic variations, spelling variants, acronyms and other sources of word variability are identified and handled to achieve text normalization. Correspondence Analysis (Benzécri, 1973) and Ward's Hierarchical Clustering are used to identify acknowledgement patterns by disciplines based on similar lexical features. Lexico 3 (Lamalle et al., 2003) is used to build a lexical table and MultiBiplot (Vicente-Villardón, 2010) for the analysis.

Results and discussion

The entire corpus is constituted by 50,710 word occurrences, of which 10,936 are different forms. The number of papers and main lexical features of the four disciplines selected are shown in table 1.

Results of Correspondence Analysis reveal differences in the lexical patterns of the disciplines selected (see figure 1). The first two axes absorbed 92.9% of the total inertia. Evolutionary Biology stands in the fourth quadrant and the support acknowledged can be defined as technical assistance and performing experimental work (Cluster 1). Economics and Statistics & Probability are characterized by words that recognize some intellectual debt which contributes to improve the quality of research, i.e. 'peer interactive communications' (PIC) (Cluster 2). Cardio & Cardiovascular System is placed in the third quadrant -close to axis 1- (Cluster 3). The implication of companies and the concern on conflict of interest's issues characterise this discipline.

Conclusions

Text mining constitutes an interesting approach for the study of acknowledgements in publications. Although only four disciplines are studied, the existence of inter-field differences in the textual pattern of the acknowledgements is confirmed. The information included in this field goes beyond the financial data and provides interesting data about sub-authorship collaboration.

Acknowledgements

This research was supported by the project CSO2008-06310 and a JAE predoctoral grant awarded by the Spanish National Research Council (CSIC).

References

Benzécri, J.P. (1973). L'analyse de Données. Vol. 2. L'analyse des correspondances. Paris: Dunod.

Costas, R., & Van Leeuwen, T.N. (2012). Approaching the "Reward Triangle": general analysis of the presence of funding acknowledgements and 'peer interactive communication' in scientific publications. *JASIST*, *63*(8), 1647-1661.

Cronin, B. (1995). The scholar's courtesy: The role of acknowledgements in the primary communication process. Los Angeles: Taylor Graham.

Giles, C.L., & Council, I.G. (2004). Who gets acknowledged: measuring scientific contributions through automatic acknowledgement indexing. *Proceedings of the National Academy of Sciences of the United States of America, 101*(51), 17599-17604.

Kassirer, J.P., & Angell, M.A. (1991). On authorship and acknowledgements. *The New England Journal of Medicine*, *325*(21), 1510-1521.

Lamalle, C., Martínez, W., Fleury, S., & Salem, A. (2003). Lexico 3.

Université de la Sorbonne nouvelle, Paris. http://www.tal.univ-paris3.fr/lexico/.

Melin, G., & Persson, O. (1996). Studying research collaboration using co-authorships. *Scientometrics*, *36*(3), 363-377.

Rigby, J. (2011). Systematic grant and funding body acknowledgement data for Publications: new dimensions and new controversies for research policy and evaluation. *Research Evaluation*, *20*(5), 365-375.

Vicente-Villardón, J.L. (2010). MultBiplot: A package for Multivariate Analysis using Biplots. Departamento de Estadística. Universidad de Salamanca.http://biplot.usal.es/ClassicalBiplot/index.html.

Table 1. Lexical features of the corpus.

Disciplines	No. Papers	Word occurrences	Word forms	Max. frequency	Hapaxes
Cardiac & Cardiovascular Systems	806	11608	2436	605	1509
Economics	624	4734	1351	344	844
Evolutionary Biology	310	23600	5104	1287	3456
Statistics & Probability	309	10767	2045	755	1315

Hapaxes: words with only one occurrence in the corpus.

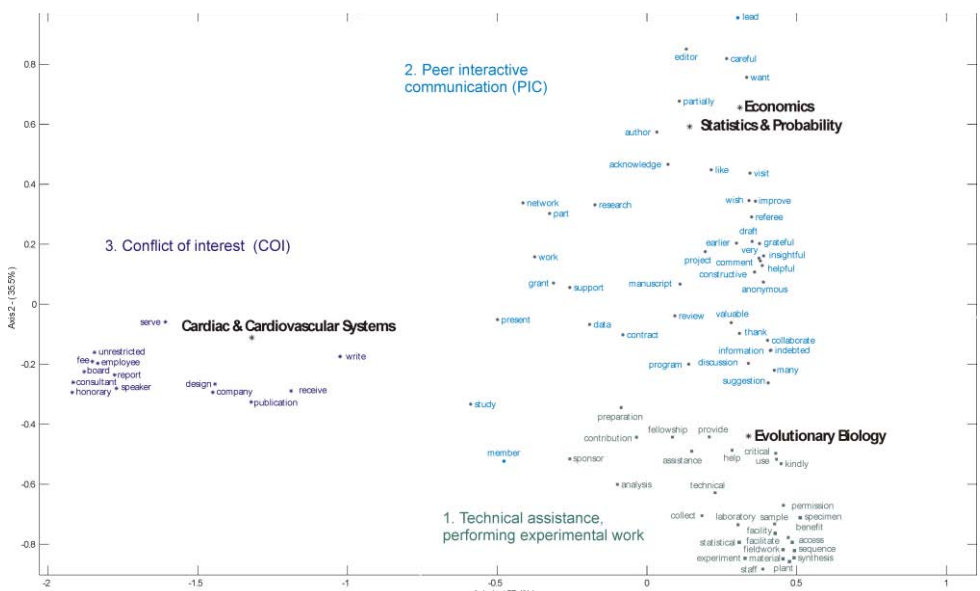


Figure 1. Correspondence Analysis of the different clusters obtained on the principal factorial plane 1-2.

REGULARITY IN THE TIME-DEPENDENT DISTRIBUTION OF THE PERCENTAGE OF UNCITED ARTICLES: AN EMPIRICAL PILOT STUDY BASED ON THE SIX JOURNALS

Zewen Hu[1,2] and Yishan Wu[1]

huzewen915@163.com; wuyishan@istic.ac.cn
[1] Institute of Scientific & Technical Information of China, No.15 Fuxing Road, 100038 Beijing (China)
[2] Department of Information Management, Nanjing University, No.22 Hankou Road, 210093 Nanjing(China)

Introduction

In the scientific community, there exists frequent uncitedness to the mediocre, the low quality, the unintelligible, the irrelevant, the valuable but undiscovered or forgotten, the par excellence, and the well known documents (Garfield, 1973). Price (1965) estimated that about 35 percent of all published papers in 1961 are not cited at all in any given year and 10 percent is never cited over a period of ten years. It is an accepted fact of academic life that some papers, in whatever discipline and wherever published, will never be cited (Burrell, 2002). Though articles that are less frequently cited in a shorter time window may be of great value that has not been found and utilized due to various reasons, if they fail to attract attention from the peers over a long period of time following their publication, then they might have weakness in relevance, importance, popularity, novelty, quality or impact. However, up to now, there is very scarce literature on the time-dependent pattern of non-citation rates of articles, and on what distribution model can fit their time-dependent pattern, as well as on the factors influencing the non-citation rate.

Data and Method

We adopt two criteria for our selection of sample journals: enough number of articles produced per year and high enough IFs in two JCR categories: Information Science and Multi-disciplinary Science, Then we draw the publication and citation data on 6 selected sample journals--*Nature, Science, Journal of the American Society for Information Science and Technology (JASIST), Scientometrics, Journal of Information Processing & Management (JIPM), and Journal of Documentation (JOD)* from *Web of Science*. After that, we use a software named "Origin 8" to draw the time-dependent scatter plots of the percentages of uncited articles in twelve different time windows from one year to twelve years following their publication in 1998 and 1999. Simultaneously, we use a three-parameter negative exponential model (Eq.1) to fit them and figure out the corresponding parameters. Finally, for exploring how great an influence the length of articles exerts on the probability that they get cited in the future, we analyse the percentage of uncited articles with different pages in a wide time window of twelve years after

their publication in 1992-1999. The expression of Eq.1 is

$$P(X_t=0)=K+Ae^{-t/s} \quad (1)$$

Here, $P(X_t=0)$ represents the percentage of uncited articles in a time window of t years after publication, $t \geq 1$. The variable "K" means the deviation value between the actual percentages of uncited articles and the expected percentages according to the fitting curve, while "A" is the amplitude of decrease for the percentage of uncited articles along with time. The variable "s" is the rate of obsolescence, which indicates the probability of uncited articles' staying in the state of uncitedness as time elapses. We call this state "sleeping" (until being awaken by the first citation) coefficient.

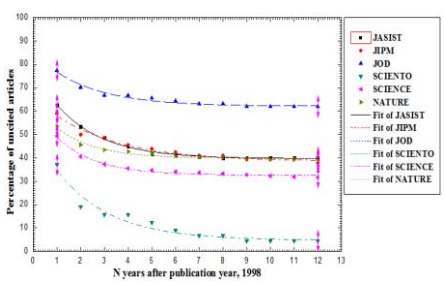


Figure 1. The time-dependent scatter plots and the fitting curves of the percentages of uncited articles after their publication in 1998.

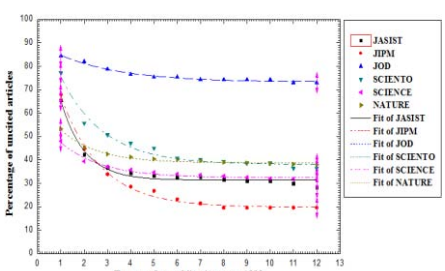


Figure 2. The time-dependent scatter plots and the fitting curves of the percentages of uncited articles after their publication in 1999.

Results

Figures 1-2 show the time-dependent scatter plots of the percentages of uncited articles in twelve different time windows from one year to twelve years following their publication, and the fitting curves with Eq.1. In these figures, the different shapes of scatter points represent different journals, while the vertical arrows represent the deviation degree between scatter plots and fitting curves.

From Figures 1-2, we can find some common patterns on the time-dependent distribution of the percentages of uncited articles. (1) The deviation between scatter plots and fitting curves is very low. (2) In the beginning shorter time window, each journal has a higher percentage of uncited articles. For example, in a time window of one year after publication, the average percentage of uncited articles in *JOD* is the highest one, reaching 80.9%, while the average percentage of uncited articles in *SCIENCE* is lowest, still reaching 48.7%. (3) As the time window becomes wider and wider, the average percentage of uncited articles in each journal begins to descend with varying degrees. For example, the average percentage of uncited articles in *JOD* drops from 80.9% in a time window of one year to 67.5% in a time window of twelve years, with a total decline of 32.6 percentiles (relative to the origin 100%), while the average percentage of uncited articles in *SCIENCE* drops from 48.7% in a time window of one year to 31.6% in a time window of twelve years, with a decline of 68.4 percentiles. (4) In a time window of twelve years after publication, the average percentage of uncited articles in each journal keeps a stable value, with very few changes. (5) The Eq. 1 can well fit the time-dependent scatter plots of the percentages of uncited articles as shown

in Figures 1-2. The average R^2 values, showing the goodness of fitting curves for each journal, are all above 97%. (6) The average amplitude value (A) of the decline of non-citation rate along with time for *JASIST* reaches the highest value of 64.5, while *JOD* keeps the lowest value of 19.9. (7) *JASIST, NATURE* and *SCIENCE* keep the lowest sleeping coefficient (S) values of 1.5, 1.6 and 1.7. While the sleeping coefficient (S) values for *SCIENTOMETRICS*, *JIPM* and *JOD* are 2, 2.3 and 2.5, respectively.

Table 1 shows the number (A) of articles with different pages published in six sample journals in 1992-1999, as well as the number (U) and the share (S) of uncited articles with different pages in a wide time window of twelve years after their publication.

Table 1. The relation between article length and the probability of getting cited

1992-1999	**1-2 pages**	**3-4 pages**	**5-6 pages**	**7-8 pages**	**9-10 pages**	**>10 pages**
A	32148	12011	3261	905	354	1292
U	18682	690	73	25	21	66
S(%)	58.1	5.7	2.2	2.8	5.9	5.1

From Table 1, we can observe the following points. (1) The number (A) of the short articles with 1-2 pages is very large. For example, there are 32148 short articles with 1-2 pages in total 49971 articles, its share in total reaches 64.3%, while there are only 5812 long articles with 5 and more pages, with a share of 11.6%. (2) The percentage of uncitedness in short articles is also very high. For example, the percentage of uncitedness in the 32148 short articles with 1-2 pages is 58.1%, while that in the 5812 long articles is only 3.18%, which may bring us to conclude that the probability of getting cited in the future for current uncited long articles is far higher than that for short articles. It seems that the scientific community is prone to cite long articles, which are more possible to provide more solid argument for their ideas. (3) The length of articles has a very great effect on the probability whether they will get cited in the future or not. For example, with the increase of article length, the share (U) of uncited articles in all journals drops from 58.1% in short articles with 1-2 pages to 5.1% in long articles with 11 and more pages, with a decline of 53 percentiles.

Acknowledgments

This study is supported by the National Natural Science Foundation of China (Grant No. 70973118) and the Innovation Plan Program of Scientific Research for College Graduate Students in Jiangsu Province (Grant No. CXZZ12_0075).

References

Garfield, E.(1973). Uncitedness III-The Importance of Not Being Cited. *Current Contents*, 8, 5-6.

Price, D. J. D. (1965). Networks of scientific papers. *Science*, 145, 510-515.

Burrell, Q.L. (2002). Will This Paper Ever Be Cited?. *Journal of the American Society for Information Science and Technology*, 53(3), 232–235.

TOPOLOGICAL TOPIC TRACKING – A COMPARATIVE ANALYSIS

Alexander Struck[1]

[1]*Alexander.Struck@ibi.hu-berlin.de*
Humboldt-Universitaet zu Berlin, Berlin School of Library and Information Science,
Unter den Linden 6, 10099 Berlin, Germany

Introduction

Topological methods for identifying and tracking community development have been proposed but not compared yet. Due to the heterogeneous nature of thresholds, viewpoints and data sets tracking results may differ significantly. This may turn into a problem if e. g. funding policy decisions are based on how well a topic 'performed' over time. The on-going research investigates several approaches and their dependencies.

Data

About 11 per cent of Thomson Reuter's WoS subject categories "Information Science & Library Science" and "Computer Science, Information Science" have been downloaded. Journals were selected if they had a 5-year Impact Factor higher than 1. This threshold has been chosen to achieve a middle-ground between a broad range of topics and their connection via citations. The resulting journal set was then expanded by the 16 most cited journals with exceptions of 'Science' and the like. This levels out the hard cuts made by the category borders. Overall 50 years of (ca. 64,000) WoS records reflecting scientific communication (e. g. 'proceedings paper') are part of the investigation.

Graph Partitioning

The Louvain algorithm (Blondel, 2008) provides clusters and hierarchy levels without threshold dependence. It has been shown that it leads to good results on benchmark tests (Lancichinetti, 2009). Similarity is indicated by bibliographic coupling and measured by Salton's Cosine index.

Tracking Approaches & Cluster Similarity

Direct Citation to and from 'core documents'

Glaenzel and Thijs (Glaenzel W. a., 2011) suggest the use of ‚core' documents and look for links from and to clusters in adjacent time periods. (Schubert, 2009) suggested „Using the h-index for assessing single publications" and „define(s) the h-index, h, of a publication as the citation h-index of the set of papers citing it [...]". (Glaenzel W. , 2012) genera-lized the ‚lobby index' for undirected graphs as: "Core nodes are nodes with at least h degrees each, where h is the h-index of the graph." This has been used to identify cores and links connecting clusters.

Combined Linkage

(Small H. , 1997) introduced a „ measure of docu-ment similarity: Combined linkage" to overcome

limitations of citation-based measures. Biblio-graphic coupling, co-citation, direct and indirect links (over a third node) are considered here. The linkage combination promises connections based on more information than currently used.

$$\boldsymbol{Coupling(A,B)} = \frac{\boldsymbol{2 * direct + indirect\ links}}{\sqrt{(\boldsymbol{paperA.links + 1}) * (\boldsymbol{paperB.links + 1})}}$$

This measure has been generalized for cluster similarity using size-relative counts. It also allows direct comparison of methods used to connect clusters over time.

Experiments

1740 clusters were compared with each other (trailing 'X' below) and also just with the adjacent one-year time frame. Several linkage options have been tested and the resulting graphs compared by noting Jaccard similarity to other graphs.

„AB1" creates graphs similar to the original since links between clusters appear if at least one direct citation exists. „ABcde1" is a graph that links A and B if at least one of the coupling methods is present. The „ABcde" graph has a link between A and B if the sum of direct citation links and papers ‚c', ‚d' or ‚e' exceeds the threshold. „AB(X)" denotes direct citation between clusters of adjacent (any) years. Co-citation coupling of A and B is present if at least that many papers ‚ci' are present in adjacent (or ‚ciX' any) years. „di(X)" graphs couple clusters bibliographically and „eiX" covers graphs of longitudinal coupling. And „core" denotes graphs that are created if at least half the noted threshold is met. This can be justified due to the fact that only a fraction of the articles are considered in linking.

Results & Discussion

Graphs resulting from the same similarity method for clusters but using either the next or all timeframes differ. This suggests that the strongest connection does not necessarily exist between adjacent time frames.

Since a lot of information is discarded in topological tracking hardly any graph compares to the maximal possible graphs of 'AB1' or 'ABcde1'. A notable exception is the sum of possible linkage (ABcde/X) which is the least restrictive threshold tested here and includes the most information. Bibliographic coupling (di/X) appears most similar since usable citation information is available outside of the data set while direct citation, co-citation and longitudinal coupling approaches are limited to the 50 years set. One could argue that the data set in conjunction with the coupling method used influences the resulting cluster strings.

Increasing the threshold creates graphs with less clusters and less edges assuming stronger connections between clusters. Here it should be noted that bibliographic coupling seems to behave more stable than using ‚core' links only. While the ‚core documents linkage' approach creates graphs very similar to ‚direct citation' (AB) this changes when comparing the threshold 10 and 50 (resp. 5 and 25 for ‚core'). At the increased threshold ‚core' is now more similar to bibliographic coupling (di/diX) and also to the less restrictive ‚ABcde(X)'. At higher threshold of direct citation links it is apparent that adjacent years (AB) do not match to longer distances for this approach (ABX). Here, the co-citation and longitudinal coupling show interesting simiarities with ABX. That suggests that certain approaches perform differently depending on thresholds like time distance.

It does make a difference if one chooses co-citation or bibliographic coupling to connect clusters. Co-citation analysis may even that out but hasn't been considered here.

The combined linkage (ABcde/X) performs stable at different thresholds and against many other approaches which may point to a consensus that will be investigated further.

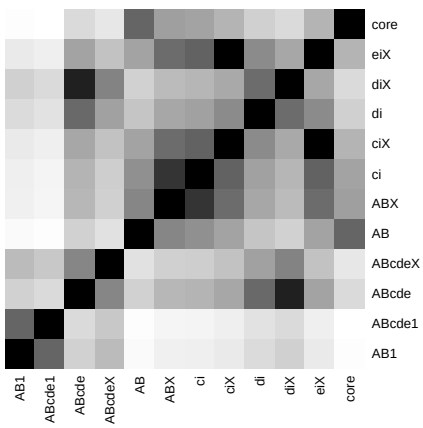


Figure 1. Threshold of 10 cluster connections

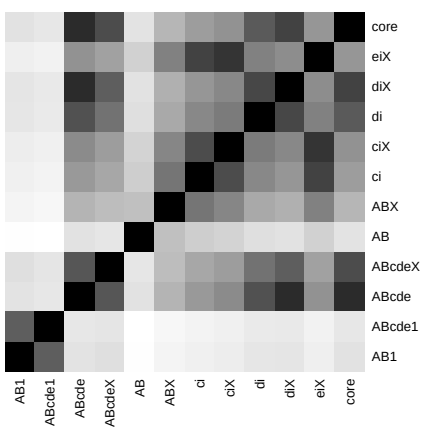


Figure 2. Threshold of 50 cluster connections

An extended version of this paper is available at: http://141.20.126.172/~div/downloads/Topological_Topic_Tracking_pt1-Struck.pdf

References

Blondel, V. a. (2008). Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment* , S. P10008.

Fortunato, S. a. (2007). Resolution limit in community detection. *Proceedings of the National Academy of Sciences* (S. 36). National Acad Sciences.

Glaenzel, W. a. (2011). Using 'core documents' for detecting new emerging topics. Proceedings of ISSI 2011-the 13th International Conference on Scientometrics and Informetrics.

Glaenzel, W. (2012). The role of core documents in bibliometric network analysis and their relation with h-type indices. *Scientometrics* .

Good, B. a. (2010). Performance of modularity maximization in practical contexts. *Physical Review E , 81* (4), S. 046106.

Lancichinetti, A. a. (2009). Community detection algorithms: A comparative analysis. *Physical Review E , 80* (5), S. 056117.

Lancichinetti, A. a. (2011). Limits of modularity maximization in community detection. *Physical Review E , 84* (6), S. 066122.

Schubert, A. (2009). Using the h-index for assessing single publications. *Scientometrics , 78* (3).

Small, H. (2006). Tracking and predicting growth areas in science. *Scientometrics , 68* (3), S. 595-610.

Small, H. (1997). Update on science mapping: Creating large document spaces. *Scientometrics , 38* (2).

TOWARDS AN AUTHOR-TOPIC-TERM-MODEL VISUALIZATION OF 100 YEARS OF GERMAN SOCIOLOGICAL SOCIETY PROCEEDINGS

Arnim Bleier and Andreas Strotmann

{arnim.bleier, andreas.strotmann}@gesis.org
GESIS - Leibniz Institute for the Social Sciences, Unter Sachsenhausen 6-8, Cologne (Germany)

Introduction

Author co-citation studies (Zhao & Strotmann, 2008) employ factor analysis to reduce high-dimensional co-citation matrices to low dimensional and possibly interpretable factors, but these studies do not use any information from the text bodies of publications. We hypothesise that term frequencies may yield useful information for scientometric analysis. In our work we ask if word features in combination with Bayesian analysis allows for well-founded science mapping studies. This work goes back to the roots of Mosteller and Wallace's (1964) statistical text analysis using word frequency features and a Bayesian inference approach, tough with different goals. To answer our research question we (i) introduce the data set on which the experiments are carried out, (ii) describe the Bayesian model employed for inference and (iii) present first results of the analysis.

The DGS Dataset

The collection of documents D we use in the experiment covers ~100 years of proceedings (from 1910 to 2006) of meetings of the *Deutsche Gesellschaft für Soziologie* (DGS), a total of 5,010 documents. Early proceedings had been scanned and OCRed, others were used in original digital form. Metadata for the documents included 3,661 distinct full names of authors J.

From each document, the 21st-320th words were extracted. After unifying word case, we removed stop-words, short and/or rare words (< 4 letters; > 10 occurrences; mostly OCR fragments) and words found in more than half of the documents, resulting in 1,067,128 occurrences from a vocabulary V with 12,665 distinct words.

Statistical Model

We now review the statistical model we employ to relate authors and documents via a flexible number of topics. Following a common notation (Rosen-Zvi, 2004), a document d is modelled as a vector of N_d words, $\boldsymbol{w}_d$, where the ith word w_{di} is chosen from the unique terms in vocabulary V. Each document d is associated with a set of authors $\boldsymbol{j}_d$ from the set of all authors, J.

Our model assumes that documents are generated in the following steps:

1. Draw a shared discrete probability distribution from a Dirichlet Process (DP) (Teh et al. 2006) with base measure H and prior concentration parameter γ as a global mixture over topics
$\tau \sim DP(H, \gamma)$
2. For author j, draw an author specific distribution over topics from the global

topic mixture τ, with prior concentration parameter α

$\theta_j \sim DP(\tau, \alpha)$

3. For each topic k, draw a topic specific distribution over vocabulary V from the symmetric Dirichlet prior β

$\phi_k \sim Dir(\beta)$

4. For all N_d words in document d, (i) draw an author indicator x from the set of authors $\boldsymbol{j}_d$ of document d; (ii) a topic indicator z from the author specific topic distribution θ_j; (iii) the observed word w itself from the respective topic

$x_{di} \sim Discrete(\boldsymbol{j}_d)$

$z_{di} \sim Discrete(\theta_{x_{di}})$

$w_{di} \sim Discrete(\phi_{z_{id}})$

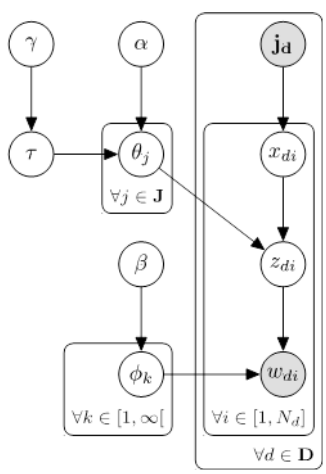


Figure 1. Non-parametric Author-Topic model.

Figure 1 illustrates the independence assumptions made by the generative storyline via plate notation. Circles represent statistical variables, with observed ones shaded. Arrows represent conditional dependence − i.e., the order in which variables are drawn. Plates indicate repetition, as indicated by universal quantifiers.

For the posterior analysis, the topic distributions ϕ_k over terms as well as the author distributions θ_j over topics are of particular interest. The former span a latent semantic space via meaningful word probabilities for each topic; the latter allow us to position each author j in this topic space.

Posterior Analysis and Visualization

The generative model is structured as a directed acyclic graph beginning from causes and ending with observed words in documents. Bayes' Rule reverses causality, and parameters of interest can be estimated from observed data and priors. We use an MCMC sampler for this posterior analysis. After running the sampler for 2,000 steps with priors γ=.5, α=.5 and β=.2, the model converged to 89 components. Samples of ϕ_k and θ_j are shown in Figure 1 and Table 1. The reader is referred to Rosen-Zvi et al. (2004), Teh et al. (2006) and Bleier (2012) for an in-depth discussion of this method.

Table 1. Most probably words for the topics.

Topic	High probability words
1	bildung, jugendlichen, schule, jugendliche, ausbildung, jugendlicher **translation**: education, school, youth
63	frauen, männern, männer, geschlecht, frauenforschung, frau **translation**: women, men, gender research
85	globalisierung, welt, globalen, grenzen, globaler, unternehmen **translation**: globalization, world, enterprize
87	europäischen, europa, integration, europäische, union, ländern **translation**: European, integration, countries

Due to space constraints we restrict the visualization in Figure 1 (using Pajek) to four of 89 components. Authors and topics are represented as square and circular nodes, resp. The size of topic nodes is proportional to their usage and the strength of the arcs proportional to

θ_{jk}, the probability for topic k specific to author j. We are not constrained to interpreting topics as only having a distinct probability for each author, but equally have for each topic k a distribution ϕ_k over distinct words in the vocabulary. Table 1 displays the six most probable terms for the sample topics of Figure 1.

Discussion

Our approach to science mapping uses a flexible version latent Dirichlet allocation to (i) identify an optimal number and set of topics for a given set of documents based on the words that occur in them, (ii) to identify the most relevant words to describe each topic, and (iii) to identify weighted links between authors and the topics of their writings. The statistical model takes into account that documents are written by multiple authors, that authors write on different topics to different degrees, and that words pertain to different topics to varying degrees.

Figure 1 shows a small fragment of a map of German sociological science based on ~100 years of DGS proceedings, inspired by the visualization of results of co-citation-based factor analysis in Zhao & Strotmann (2008), but generated fully automatically from the results of applying this statistical analysis technique to full texts.

While a full evaluation remains to be done, these results show some promise for the application of these methods in scientometric studies.

References

Bleier, A. (2012). A simple non-parametric Topic Mixture for Authors and Documents. Pre-print arXiv:1211.6248 [cs.LG].

Mosteller, F., & Wallace, D. (1964). *Inference and Disputed Authorship: The Federalists.* Addison-Wesley.

Rosen-Zvi, M., Griffiths, T., Steyvers, M., & Smyth, P. (2004). The author-topic model for authors and documents. *Proceedings of the 20th Conference on Uncertainty in artificial intelligence (UAI 04)* (pp. 487-494).

Teh, Y.W., Jordan, M.I., Beal, M.J., & Blei, D.M. (2006). Hierarchical dirichlet processes. *Journal of the American Statistical Association*, 101, 1566-1581.

Zhao, D., & Strotmann, A. (2008). Information Science during the first decade of the Web: An enriched author co-citation analysis. *Journal of the American Society for Information Science and Technology*, 59(6), 916-937

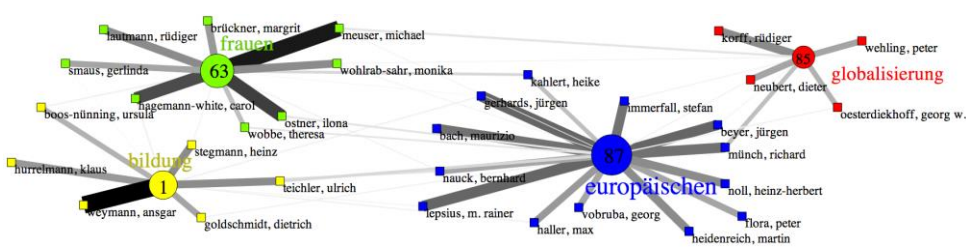


Figure 1: Excerpt of author topic analysis result: visualization of four topics and their main authors

USE FREQUENCIES OF NOMINALIZATIONS IN SCIENTIFIC WRITING IN BRAZILIAN PORTUGUESE LANGUAGE AS POLITENESS STRATEGIES AND THEIR INDEX ROLE IN THE SUBJECT INDEXING

Vânia Lisboa da Silveira Guedes[1]; Maria de Fátima Sousa de Oliveira Barbosa[2]; Maria José Veloso da Costa Santos[3] and Maria Cecília de Magalhães Mollica[4]

[1]*vanialisboa@facc.ufrj.br;* [2]*fatimma.barbosa@gmail.com;* [3]*msantos1402@gmail.com;* [4]*ceciliamollica@terra.com.br*
Federal University of Rio de Janeiro, Av. Pasteur, 250 – Urca CEP 22290-902, Rio de Janeiro-RJ – Brazil

Introduction

This study is part of "Scientific and Technological Knowledge Organization Project" and analyses the scientific writing in Chemistry subareas in Brazilian Portuguese in order to investigate the regularity in the use frequencies of deverbal nominalizations as politeness strategies in the scientific communication and their index role in the subject indexing. The research is situated in the border of the Linguistics with Information Science and aims to contribute to the understanding of the scientific writing with communication purposes. The research question is how a systematic analysis of deverbal nominalizations in scientific writing of papers can contribute to the semi-automatic subject indexing based on Bibliometrics, making information retrieval electronic systems more precise, intelligent and scientifically established.

Objectives

Central Objective

The central aim is to develop the linguistic and bibliometrics comparative analysis of the scientific writing in the Chemistry subareas: Pharmacology Industry and Food Technology.

Specific Objectives

(a) to analyze the scientific writing based on quantitative models used in semiautomatic indexing within Bibliometrics;
(b) to investigate the regularity in the use frequencies of the nominalization in ***–ção, -mento;***
(c) to contribute to the theoretical and practical approaches in the knowledge fields in discussions.
(d) to strengthen the interface of the Discourse Critical Analysis in Linguistics with Scientific Communication and Bibliometrics in Information Science.

Hypothesis

The hypothesis is that the use frequencies of nominalization

represented by *[X] v* → *[[X] v* -ção*] N* is predominant in the semantics fields in analysis. So it is present in the scientific writing of these papers as index terms and politeness strategies.

Theoretical framework

The theoretical framework used was the subject indexing, Zipf's Laws and Goffman Transition Point (Pao, 1978) in the Bibliometrics, as well as the genre analysis theory (Bazerman, 2006; Hyland, 2009; Swales, 1990), the Lexical Theory (Basílio, 2007; Chomsky, 1970) and Critical Discourse Analysis (Eggins, 2004; Van Djck, 2012) within Linguistics. Bibliometrics is the science that presents a set of empirical principles based on mathematical and statistical methods to investigate, assess and quantify the written communication processes. The Bibliometrics analysis establishes relevant indicators in a knowledge field highlighting the quantitative aspects of production, dissemination and use of scientific information. Among the more used bibliometrics laws there are the Zipf's Laws used for subject indexing related to words occurrence frequencies in a given text, enriched by Goffman Transition (T) Point, a method of selecting index terms directly suggested by Goffman (apud Pao, 1978). This method indicates a region from the list of words used in a scientific text with the highest semantics content. Goffman T Point is represented mathematically by the expression above, where *n* represents the Point T; I_1 is the number of words that has a frequency 1.

$$n = \frac{-1 + \sqrt{1 + 8I_1}}{2}$$

In Linguistics, the deverbal nominalization, (Basilio, 2007), refers to the set of processes that form nouns from verbs. The author explains that the nominalization contains aspects syntactic and semantic textual and play functions of designation of process, action, state etc. Basilio (2007), Swales (1990), Hyland (2009), Eggins (2004) and other authors emphasize that the use of nominalizations characterize heavily the scientific discourse. Guedes (2010) demonstrates the importance of the analysis of the deverbal nominalizations in scientific writing about Viniculture for the subject indexing in Information Science. Santos (2009) verifies the possibility of applying the Zipf's Laws and the Goffman T Point in a scientist personal archive in Zoological area in order to find words with a high semantic content for subject indexing. Barbosa (2010) analyses the linguistic processes used to represent politeness strategies in messages exchanged among the students and the mediators of the Distance Learning Courses. Guedes, Barbosa and Santos (2012) analyses the productivity and the recurrence of nominalized structures with of index terms function and politeness strategies in the scientific writing with communication purposes.

Methodology

The methodology consisted of the following steps:

(1) sample definition - The relevant periodical titles were selected in the QUALIS evaluation system of the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES). Therefore, two papers about Pharmacology Industry and two papers about Food Technology were selected in the Scientific Electronic Library Online (Scielo).

(2) count of words - these papers were processed by *Software Rank Words 2.0;*

(3) listing and ranking of words - the software produced a table in 3 columns, distributed as follows: words, order of decreasing frequency and rank of words;

(4) applying Zipf´s Laws and the formula of Goffman T Point - it was identified the frequency where the transition from high frequency words for low frequency occurs;
(5) defining Goffman T region – it was identified in the ranking the region that concentrates the most content-bearing words that should be used for indexing and provide greater precision in information retrieval systems;
(6) verification of nominalizations recurrence in –**ção** and –**mento**, within Goffman T region, and their index terms functions, as well as the politeness strategies in scientific communication.

Results

The table 1 shows the concentration region of words and the nominalizations use frequencies that represents the politeness strategies and index terms in the text number one.

Table 1 – Concentration Region

Rank	Word	Frequency
29	processo	20
35	cáries	15
36	**produção**	14
38	bactérias	14
40	**crescimento**	13
43	produtos	12
44	**concentração**	12
45	quantidade	11
47	bucal	11
49	Streptococcus	10

Conclusion

The results point to recurrence of nominalizations in **–ção** with high semantic content in Goffman T Region and confirmed the established hypothesis and the importance of theoretical and descriptive approaches of deverbal nominalization for the subject indexing within information research systems. Finally, the recurrent pattern in **-ção** contributes to the knowledge of lexical-morphological features more productive in texts, making it of great importance to the process of identifying the scientific information content.

References

Barbosa, M.F. (2010). *(Im)polidez em EAD.* Tese (Doutorado em Linguística) - UFRJ, Rio de Janeiro.

Basílio, M. M. de P. (2007). *Teoria lexical*. 8. ed. revista e atualizada. São Paulo: Ática.

Bazerman, C. (2006). *Gêneros textuais, tipificação e interação*. São Paulo: Editora Cortez.

Chomsky, N. (1970). Remarks on nominalization. In: R. Jacobs & P. Rosenbaum (Org). *Readings in English transformational grammar*. Waltham, MA: Ginn.

Eggins, S. (2004). *An introduction to systemic functional linguistics*. 2. ed. Nova York: Continuum International Publishing Group.

Guedes, V. L. da S. (2010). *Nominalizações deverbais em artigos científicos*: uma contribuição para a análise e a indexação temática da informação. Tese (Doutorado em Linguística) – UFRJ, Rio de Janeiro.

Guedes, V.; Barbosa, M.F.; Santos, M.J. (2012) Recorrência de nominalizações deverbais em resumos de cartas científicas em língua portuguesa como estratégia linguística de polidez. Lisboa, Portugal: Universidade Aberta.

Hyland, K. (2009) *Academic discourse*: English in a global context. New York: Continuum International Publishing Group.

Pao, M. L. (1978). Automatic text analysis based on transition phenomena of word occurrences, *JASIS*, 29, p. 121-124.

Santos, M. J. V. da C.(2009) Correspondência científica de Bertha Lutz: um estudo de aplicação da lei de Zipf e Ponto de transição de Goffman em um arquivo pessoal. *Ponto de Acesso*, .3, p.317-323.

Swales, J. M. (1990). *Genre analysis*: English in academic and research settings. Cambridge: Cambridge University Press.

Van Djck (2012). *Discurso e contexto*: uma abordagem sociocognitiva. Trad . R. Ilari. São Paulo: Contexto

Zipf, G.K. (1949) *Human behavior and the Principle of Least Effort*. Cambridge, MA: Addison-Wesley.

A VISUALIZATION TOOL FOR TOPIC EVOLUTION AMONG RESEARCH FIELDS

Kun Lu[1], Wei Lu[2], Qikai Cheng[3], Shengwei Lei[4] and Jiangen He[5]

[1]kunlu@whu.edu.cn, [2]reedwhu@gmail.com, [3]chengqikai0806@gmail.com, [4]leoray1989@gmail.com, [5]hejg@vip.qq.com
School of Information Management, Wuhan University, Luojiashan Road No.1, Wuhan, Hubei, China, 430072

Introduction

Large amount of scientific productions have posed big challenges to understand the underlying structure. Visualization tools can help to unveil the intriguing features of the structure in a rather intuitive way. Depending on the use of the visualization tool, it may help to delineate the research areas within a field, outline the research interests of a specific author, or present other structures of interests. It is common that a research topic originated in one field spreads to other related fields and evolves over time. However, few existing tools are specially designed to visualize the topic evolution among research fields. The purpose of the current study is to propose a new visualization tool that focuses on unveiling the topic evolution among research fields over time.

Proposed method

Overall Structure

The purpose of our tool is to visualize the evolution of a research topic among different fields. A research topic could be related to multiple fields so the data is multi-dimensional. Temporal data is needed to describe the evolution. We proposed a new method to visualize the compound of the types and named it VIS-TopicEvo which is short for "Visualization for Topic Evolution". A summary of the overall structure of VIS-TopicEvo is provided in Figure 1 where *C* represents a research field (or a reference point), *R* represents a research topic (or a data point), *H* denotes the hierarchical structure of the research fields, *T* represents the temporal data and *M* denotes the multi-dimensional relationship between a data point and research fields. The G_r, which consists of R_1, R_2,…,R_5, reflects the evolution of a research topic over time. The G_H reflects the hierarchical structure of research fields. And the G_M represents the multi-dimensional relationship between the research topic and related fields.

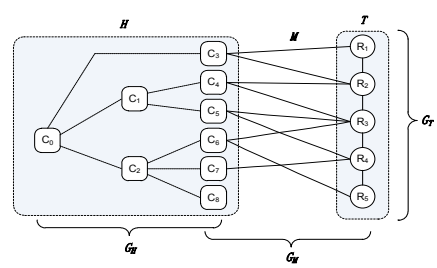


Figure 1.The overview of data structure in VISTMH

Multi-dimensional Data Projection

In VIS-TopicEvo, a data point (i.e. a research topic at a given time slot) is related to multiple reference points (i.e. research fields). The multi-dimensional data needs to be projected onto a 2D or

3D visual space. We use a similar method as in Olsen (1993): an n-dimensional space is represented as a regular polygon with n vertices. Each vertex serves as a reference point. And data points are plotted as circle icons within the polygon according to their relatedness with the references points. We use the set $C=\{C_i(c_1,c_2,...,c_i,...,c_p)\}(i = 1,\ldots,p)$ to denote the reference points. For each reference point $C_i(c_1,c_2,...,c_i,...,c_p)$, we define $c_i=1$ and all the other elements $c_1,\ldots,c_{i-1},c_{i+1},\ldots,c_p=0$. To construct a regular polygon, the position of C_i is determined by the coordinate $V_i(x_i, y_i)$ that is calculated as follows:

$$P(C_i) = V_i(\cos(\frac{2\pi i}{p}) \cdot radius, \sin(\frac{2\pi i}{p}) \cdot radius) \quad (1)$$

where the *radius* is a parameter to adjust the size of the picture. A data point R_j is defined as $R_j=(r_1,r_2,\ldots,r_i,\ldots,r_p)$ where r_i represents the relatedness score between the data point and the ith reference point C_i. We define r_i as follows:

$$r_i = \frac{\sum_{i=1}^{p} c_i * f_i}{(\sum_{i=1}^{p} c_i^2 * \sum_{i=1}^{p} f_i^2)^{1/2}} = \frac{f_i}{(\sum_{i=1}^{p} f_i^2)^{1/2}} \quad (2)$$

where c_i is the ith elements of the ith reference point, f_i is the number of times a research topic co-occurs with the ith reference point. The co-occurrence of a research topic and a reference point (i.e. a research field) is defined as an observation of a published research paper that covers the research topic and is categorized into the research field. In this way, we projected our multi-dimensional data onto a two dimensional visual space where the reference points (i.e. research fields) are vertices and the data points (i.e. a research topic at different time slots) are within the regular polygon according to their relatedness with the reference points. If they are plotted close to a reference point that indicates they are closely related with that research field at the given time slots, and vice versa.

Topic Evolution

The G_T component in Figure 1 incorporates the temporal data into VIS-TopicEvo. It consists of a sequence of data points that represent the evolution of a research topic over time. To trace the evolution visually, we plotted the data points at different time slots in different colours. Then we use cubic spline interpolation (Schumaker, 2007) to construct a smooth curve to connect the data points. An example is given in Figure 2 to illustrate our approach to visualize the topic evolution.

Figure 2. An example of topic evolution over time

The four vertices of the rectangle are the reference points which represent four research fields. The colour bar at the bottom indicates the temporal information of the data points. And the sequence of the data points unveils the evolution of the relationship between the research topic and the four fields over time.

Results

In this section, we will present our initial results. We first manually identified four hot interdisciplinary research topics from the author keyword field (i.e. "DE" field) of the bibliographic records of the publications

in the top journals of the "information science & library science" category defined by Journal Citation Report 2009 (the same journals as in Lu & Wolfram, 2012). The topic we selected is "information retrieval". We then conducted subject searches through WoS portal using the topics mentioned above and downloaded the results. Only two types of documents are kept in the data collection: articles and conference proceedings. Other types are less likely to reflect the original research contributions. The topics, queries and the time ranges we used are summarized in Table 1:

Table 1. The queries and time range used to collect our data

Topic	*Query*	*Time range*
information retrieval	"information retrieval"	1956-2011

To obtain the co-occurrence data (i.e. f_i in equation 2) of the research topics and the research fields, we counted the number of times a given WoS category (i.e. "WC" field) appears in these bibliographic records for each topic. The more frequently a given category appears in the records, the closer relationship the topic has with the category. It is possible for a publication to be assigned to more than one categories. In this case, we added the counts to all of them equally.

We deliberately selected a longer time span to trace the development of the topic after Mooers first coined the term in 1950. However, we did not find any publications from WoS in our search results during the time period of 1950 to 1955. So we eventually set the time range from 1956 to 2011. As the time range is longer, to avoid too many data points that overwhelm the picture we adjusted the time interval between the neighbouring data points to four years instead of using the one year interval as in other pictures. We can see from the Figure 3 that the topic of "information retrieval" has a closer relationship with the field of Information Science & Library Science (LIS) in the early years, and then moves towards the field of Computer Science gradually. From the beginning of this century, Computer Science became the dominant contributor to this topic. However, we can also find that the curve has a tendency to move back a little bit to LIS field more recently.

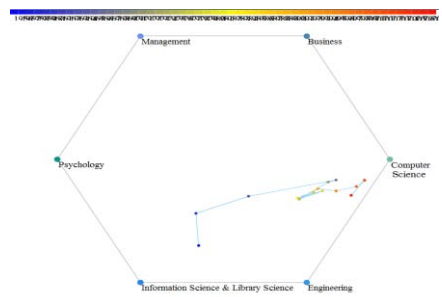


Figure 3.Topic evolution of "information retrieval"

Conclusion

Many visualization tools have been developed to help understand the structure of science. While few existing tools are specially designed to visualize the topic evolution of an interdisciplinary research topic among related research fields. We designed VIS-TopicEvo to address this particular problem. Our initial results show the promise to trace the development and evolution of a research topic among related research fields visually. The pictures provide an intuitive way to understand the trace of a topic and help to gain a historical perspective on it.

References

Lu, K., & Wolfram, D. (2012).Measuring author research relatedness: A comparison of word-based, topic-based, and author cocitation approaches. *Journal of the American Society for Information Science and Technology*, 63, 1973-1986.

Olsen, K. A., Korfhage, R. R., Sochats, K. M., Spring, M. B.& Williams, J. G. (1993). Visualization of a document collection: The vibe system. *Information Processing & Management*, 29(1), 69-81.

Schumaker, L. (2007). *Spline Functions: Basic Theory*. London: Cambridge University Press.

VISUALIZING THE RESEARCH DOMAIN ON SCIENTOMETRICS (1978-2012)

Qiu JunPing[1] and Lv Hong[2]

[1] *jpqiu@whu.edu.cn*
Wuhan University, Research Centre of Science Evaluation, 430072 Wuhan (P. R. China)

[2] *lyu@whu.edu.cn*
Wuhan University, School of Information Management, 430072 Wuhan (P. R. China)

Introduction

Visualization method has become an important research method, and has been used in many fields. By use this method, the scope of and trends in investigations, to some extent, their research level can be objectively evaluated. We examined the research status quo of this discipline by visualization of knowledge map on scientometrics in order to learn about its development and evolution process.

Materials and methods

The data contains all types of documents published in *Scientometrics* journal from 1978 to 2012. A full bibliographic record in Web of Science is used, each record contains fields such as author, title, abstract, keywords, and references and so on. The retrieval was finally updated on November 14, 2012 (journal *Scientometrics* was finally updated on October 2012). The resultant dataset contains a total of 3,324 records, 1679 (50.51%) of total set which were published during 2002-2012. By change of the number of papers reflects the study of scientometrics is focus. Information visualization tool with Java application named CiteSpace (Chen, 2006) was selected as the research tool to assist analysis, and version is 3.0.R5.

Table 1. The top 30 most highly cited scholars in the author co-citation network

No.	Scholar	Frequency
1	GARFIELD E	753
2	PRICE DJD	543
3	GLANZEL W	489
4	NARIN F	435
5	MOED HF	409
6	SCHUBERT A	395
7	Braun T	358
8	LEYDESDORFF L	357
9	SMALL H	342
10	Van Raan AFJ	316
11	EGGHE L	313
12	Hirsch JE	193
13	ROUSSEAU R	174
14	CALLON M	156
15	VINKLER P	155
16	Katz JS	145
17	Cronin B	139
18	MERTON RK	135
19	MARTIN BR	133
20	COLE S	130
21	Bornmann L	129
22	Zitt M	128
23	White HD	125
24	LUUKKONEN T	123
25	MORAVCSIK MJ	120
26	COLE JR	119
27	Meyer M	119
28	TIJSSEN RJW	117
29	NEDERHOF AJ	115
30	FRAME JD	112

Results

Analysis of most highly cited scholars

Author co-citation analysis (ACA) approach provides a useful tool for identifying major specialties and researchers and their interrelationships (Zhao & Strotmann, 2011). In terms of citations, ACA aims to provide a useful glimpse of the dynamic intellectual structure of the contributing research community. Generally speaking, authors with most citations tend to be those researchers carrying out the fundamental research tasks in their subject. Who have made important and fundamental impact on the development and evolution of scientometrics? By analysis of the author co-citation network, these scholars can be found. Tab. 1 lists the top 30 most highly cited scholars in the ACA network in the field of scientometrics.

Analysis of document co-citation network

CiteSpace represents the literature in terms of a network synthesized from a series of individual networks, and integrates these individual networks and form s an overview of how a scientific field has been evolving over time (Chen et al., 2012). The most cited articles are usually regarded as the landmarks due to their ground-breaking contributions. Top 10 most highly cited papers ranked 10th in Tab. 2, and summarizes their citation frequency, betweenness centrality and source. The table shows the top 10 most highly cited papers in the visualization map are Hirsch's (2005) paper, Price's (1963) paper, Lotka's (1926) paper, Garfield's (1972) paper, Small's (1973) paper, Schubert's (1986) paper, Price's (1965) paper, Garfield's (1979c) paper, Schubert's (1989) paper, and Katz's (1997) paper.

Table 2. The top 10 most cited papers in the scientometrics dataset from 1978 to 2012

No.	Fre-quency	Between-ness centrality	Cited refer-ence	Source (Abbreviations)
1	189	0.00	Hirsch (2005)	P NATL ACAD SCI USA
2	152	0.22	Price (1963)	NY: Columbia University Press
3	115	0.16	Lotka (1926)	J WASHINGTO N ACADEMY
4	99	0.15	Garfield (1972)	SCIENCE
5	99	0.28	Small (1973)	J AM SOC INFORM SCI
6	94	0.35	Schubert (1986)	SCIENTOMET RICS
7	93	0.24	Price (1965)	SCIENCE
8	91	0.16	Garfield (1979c)	NY: John Wiley and Sons
9	91	0.25	Schubert (1989)	SCIENTOMET RICS
10	82	0.07	Katz (1997)	RES POLICY

Analysis of research hotspots

Generally speaking, keywords are the primary content of the article extracted by paper authors, and noun phrases with high frequency from extracted from titles and abstracts are important to a paper, thus through analysis of terms such as keywords and noun phrases can help us identify hot topics of research on scientometrics. Tab. 3 lists the top 26 terms with co-occurrence frequency of over 70 times. Firstly, high-frequency terms on scientometrics indicators and models mainly include indicators with 252 times, impact with 206 times, impact factor with 137 times, h-index with 80 times, index with 76 times, and model with 58 times and so on. Secondly, high-frequency terms on science communication tools mainly include journals with 170 times, science citation index with 149 times, publications and scientific journals with 72 times respectively, publication with

66 times, scientific production with 65 times, and scientific output with 64 times and so on. Thirdly, high-frequency terms on bibliometrics include bibliometric analysis with 128 times, citation with 118 times, bibliometrics with 108 times, citations and citation analysis with 102 times respectively, and bibliometric indicators with 73 times and so on. Fourthly, High-frequency terms on include technology with 114 times, collaboration with 101 times, innovation with 90 times, performance with 80 times, research performance with 79 times and international collaboration with 75 times and so on.

Table 3. Terms with frequency more than 70 times in the scientometrics dataset

No.	Frequency	Keywords\noun phrases
1	548	science
2	252	indicators
3	206	impact
4	170	journals
5	149	science citation index
6	137	impact factor
7	128	bibliometric analysis
8	118	citation
9	114	technology
10	108	bibliometrics
11	102	citations
12	102	citation analysis
13	101	collaboration
14	99	patterns
15	92	citation analysis
16	90	innovation
17	80	h-index
18	80	social sciences
19	80	performance
20	79	research performance
21	76	index
22	75	international collaboration
23	73	bibliometric indicators
24	73	research performance
25	72	publications
26	72	scientific journals

References

Chen, C. M. (2006). CiteSpace II: Detecting and visualizing emerging trends and transient patterns in scientific literature. *Journal of the American Society for Information Science and Technology*, 57(3), 359-377.

Zhao, D. Z. & Strotmann, A. (2011). Intellectual structure of stem cell research: a comprehensive author co-citation analysis of a highly collaborative and multidisciplinary field. *Scientometrics*, 87(1), 115-131.

Chen, C. M., Hu, Z. G., Liu, S. B. & Tseng, H. (2012). Emerging trends in regenerative medicine: a scientometric analysis in CiteSpace. *Expert Opinion on Biological Therapy*, 12(5), 593-608.

WEB 2.0 TOOLS FOR NETWORK MANAGEMENT AND PATENT ANALYSIS FOR HEALTH PUBLIC

Jorge L Magalhães[1] and Luc Quoniam[2]

[1] *jorgemagalhaes@far.fiocruz.br*
Oswaldo Cruz Foundation/Fiocruz, CEP 21041-250, Rio de Janeiro; Capes Fellow 12.298-3 (Brazil); Aix-Marseille Université (France)

[2] *mail@quoniam.info*
Université du Sud Toulon-Var; Université Paul Cézanne Aix-Marseille III (France)

Introduction

According to the Organization for Economic Cooperation and Development (OECD), 55% of global wealth is in the knowledge (OECD, 2008). In the same way Drucker (Drucker, 2006) points out that increase in the knowledge generation will occur with the increase of the knowledge management. New trends are influencing industrial development of countries like knowledge used as main resource and the learning as a central process. In this sense, it is essential always to broaden expertise base in human resources and hence increasing the innovation potential (Lastres, HMM & Sarita, A, 1999).

Forming competencies to innovate requires previously thinking as an intelligence cooperative can transforming the knowledge construction in collaboration with peers at work. This mindset requires collaborative development processes capable of producing high quality information for scientific and technological knowledge. In this scenario, experts have unrestricted access to information created by the scientific community, collaborative review of the contributions of members, governance based more on authority than on sanctions and involvement in integrated levels and responsibilities (Ambrosi, A, Peugeot, V, & Pimenta, D, 2005). Moreover, within an interactive environment requires management tools to aid in decision making (Vincent & Singer, 2010). So, this work aimed to evince how Web 2.0 tools can help developing and undeveloped nations with network management and patent analysis for health care such as tuberculosis – a global threat (figure 1).

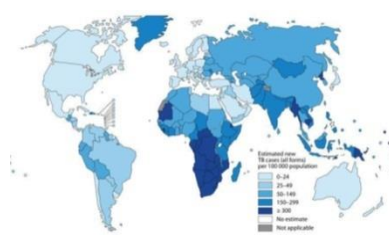


Figure 1. Estimated TB incidence rates (WHO, 2011).

Thus, is important to emphasize that intensity in research drugs and medicines has contributed significantly for improve innovation and technological development in the country's health. Likewise, co-relation of a lot of experts in the world with several knowledge bases can contribute to generation of new approaches and results as well as assist in better decision

making by managers of companies, governments and organisations.

Methodological Procedures

Use of survey databases indexed with data analysis and correlations via Web 2.0. It was identified and analysed countries and publications across research networks in tuberculosis worldwide. Moreover, it is demonstrated also a specific example of technological innovation management using tuberculosis patents.

Results and Discussion

Knowledgement management

On grounds need to better management of information from the "Knowledge Age" should be considered an adaptation to actual conditions of each local culture and collaboration for R,D&I through collaborative networks for dissemination of the knowledge aiming development and innovation. (Le Moigne, Jean-Louis, 1994; Quoniam, L, Lucien, A, 2010). Like this process is increasingly complex it becomes need multidisciplinary teams for a systemic view given that search engines have evolved from manual information's for portals or websites dedicated (Web 1.0 to Web 2.0) - passed for a massive amount of information in an automated model (Quoniam, L, Lucien, A, 2010).
Thus, considering the democratization of knowledge provided by the Web 2.0 tools using open access, it is possible to demonstrate the use of indicators for non-specialists democratized.

Public health matter

WHO Constitution enshrines as a fundamental right to health of every human being access to timely, acceptable, and affordable health care of appropriate quality ("WHO | The right to health", 2012). However, this "health" does not reached properly to most of the world population (WHO, 2008) of which 80% live in middle or low income countries. Its worst when there is a lack of medicines for neglected diseases (ND) which affect mainly populations with low purchasing power – they do not provide sufficient incentive for pharmaceutical industry to invest in R,D&I. WHO estimates that there are about 1 billion people suffering some ND such as Tuberculosis (TB). Situation is best understood when it's think in neglected populations, i.e., include not only new treatments for ND but also access to antimicrobials, affordable medicines for diseases with global impact as diabetes and cancer (Moon, Bermudez, & 't Hoen, 2012). Figure 2 shows a scenario of global disease alert, among them TB.

Figure 2. Global health map – Dez/2012.

WHO estimates that it affects two billion people, which means that a third of the world's population is infected with the bacillus Mycobacterium tuberculosis (M.tb). TB has no new drug for more than half a century.

Information technology

Mapping and location partnerships encourage better planning of R,D&I for businesses and institutions. It's possible to analyse expert information about their work and performance. Info planned and organized provides subsidies managers to define public policies and stimulate research. Likewise, in management of DN field whether in prevention, control, treatments and new technologies

(Magalhaes, JL, Antunes, AMS, & Boechat, N, 2012). Morel et al (2009) shows that co-authorship network analysis could become an important tool for international organizations or partnerships targeting the elimination or diseases eradication (Morel, Serruya, Penna, & Guimarães, 2009).
Bibliometric analysis on dez/2012 in PubMed database shows double the growth compared to publications in the early 21st century, as well as relations and co-relations on theme, trends etc – about 190.000. Figure 3 shows local R,D&I in TB. There is research intensity in the countries with the highest amount of points on the map.

Figure 3. Network for TB research.

Final consideration

- The use Web 2.0 tools for analyse R,D&I in technological forecasting for TB is effective. Notwithstanding, location mapping of network for promote knowledge management in institutions.
- The democratization of the indicators serve as tools for decision makers, especially for health care in least developed countries that do not have access to new technologies.

References

Ambrosi, A, Peugeot, V, & Pimenta, D. (2005). *Enjeux de mots - Regards multiculturels sur les sociétés de l'information*. France: C&F Editions. Recuperado de http://www.decitre.fr/livres/enjeux-de-mots-9782915825039.html#table_of_content

Drucker, P. (2006). *Desafios Gerenciais para o Século XXI*. Cengage Learning Editores.

Lastres, HMM, & Sarita, A. (1999). *Informação e Globalização na Era do Conhecimento*. Rio de Janeiro: Editora Campus Ltda.

Le Moigne, Jean-Louis. (1994). *La théorie du Système Général: théorie de la modélisation*. France.

Magalhaes, JL, Antunes, AMS, & Boechat, N. (2012). *Technological Trends in the Pharmaceutical Industry: the matter of neglected tropical diseases – An overview of the Research, Development & Innovation in Brazil*. Synergia Editora. Recuperado de http://www.livrariasynergia.com.br/livros/M39700/9788561325732/tendencias-tecnologicas-no-setor-farmaceutico-a-questao-das-doencas-tropicais-negligenciadas-edicao-bilingue.html

Moon, S., Bermudez, J., & 't Hoen, E. (2012). Innovation and Access to Medicines for Neglected Populations: Could a Treaty Address a Broken Pharmaceutical R&D System? *PLoS Med*, *9*(5), e1001218. doi:10.1371/journal.pmed.1001218

Morel, C. M., Serruya, S. J., Penna, G. O., & Guimarães, R. (2009). Co-authorship Network Analysis: A Powerful Tool for Strategic Planning of Research, Development and Capacity Building Programs on Neglected Diseases. (M. Tanner, Org.)*PLoS Neglected Tropical Diseases*, *3*(8), e501. doi:10.1371/journal.pntd.0000501

OECD. (2008). *OECD Annual Report 2008* (p. 118). Paris, France: Organisation for Economic Co-operation and Development. Recuperado de

http://www.oecd.org/newsroom/40556222.pdf
Quoniam, L, Lucien, A. (2010). *Intelligence compétitive 2.0 : organisation, innovation et territoire*. France: Librairie Lavoisier. Recuperado de http://www.lavoisier.fr/livre/notice.asp?ouvrage=2139418&pos=8
Vincent, J.-L., & Singer, M. (2010). Critical care: advances and future perspectives. *The Lancet*, *376*(9749), 1354–1361. doi:10.1016/S0140-6736(10)60575-2
WHO. (2008). WHO | Sixty-first World Health Assembly. Recuperado de http://www.who.int/mediacentre/events/2008/wha61/en/index.html
WHO | The right to health. (2012). *WHO*. Recuperado 18 de janeiro de 2013, de http://www.who.int/mediacentre/factsheets/fs323/en/

Acknowledgments

Thank CAPES Grant and infrastructure support at *Aix-Marseille Université, France*, as well as METICA's Lab in *Faculté des Sciences Saint Jerôme*.

WEIGHTING CO-CITATION PROXIMITY BASED ON CITATION CONTEXT

Bo Wang[1] , Kun Ding[1], Shengbo Liu[1]

[1] *bowang1121@gmail.com, dingk@dlut.edu.cn , liushengbo1121@gmail.com*
Wiselab Dalian University of technology, No.2 Linggong Road, Ganjingzi district, Dalian, 116024 (China)

Introduction

Co-citation is defined as a linkage between two documents concurrently cited by another document (Small, 1973). Traditional co-citation analysis does not take into account the proximity of references co-cited by an article. Some references are cited within the same sentence, whereas other references may be cited in further-apart positions in an article. Many studies have confirmed that the co-citation proximity could affect the co-citation analysis (Elkiss el al, 2008; Gipp and Beel, 2009 ; Callahan et al , 2010; Liu and Chen, 2011). Some of them tried to set the weight of different co-citation proximities based on co-citation position or distance. The nearest co-citation distance was given the highest weight, and the furthest co-citation distance was given the lowest weight. The setting of the weight values was always depending on subjective experiences. They set the smaller co-citation proximity with higher weight, because people usually think that papers with smaller co-citation proximity tend to cited by the similar topic. But in different subject areas, the influence of co-citation proximity may be different. It is imprecise to use a subjective weight of co-citation proximity in various fields. How could we set the weight of co-citation proximity that could be suitable to all the fields? In this paper, we propose a weight setting method based on the similarity of the citation context of the cited papers. The citation context of a given reference can be defined as the sentences that contain a citation of the reference. For instance, the sentence "This comparison is made using BLASTX [18]" is the citation context of reference [18].

Data and Method

The co-citation proximity analysis requires not only bibliographic information, but also the full text of an article. In this research, we utilize the PubMed Central database. In particular, references and full text information from *BMC Bioinformatics*, *BMC Systems Biology*, and *BMC Biology* are extracted and analysed. The numbers of articles in the three journals are 5412, 905 and 638, respectively, from periods of 2001-2012, 2007-2012 and 2003-2012.

Co-citation proximity

Co-citations in a citing paper are considered at four levels of proximity, namely, the article level, the section level, the paragraph level and the sentence level. If two references are cited within the same sentence, the co-citation instance is called a sentence-level co-citation. If two references are cited in different sentences but within the same paragraph, it is called a paragraph-level co-citation. Similarly, two references cited in different

paragraphs but within the same section define a section-level co-citation. Finally, if two references are cited in different sections but within the same paper, we have an article-level co-citation.

The weight of different co-citation proximity

The similarity of the co-cited contexts inevitable exist some differences when they occurred in different co-citation levels. If two citation contexts contain a same topic word or more, they will be marked as similar. If they do not have same topic words, they will be marked as not similar. The similarity of the co-cited contexts was calculated by using the following formula.

$$Similarity(C_A, C_B) = \begin{cases} 1, & C_A \text{ and } C_B \text{ contain the same topic words} \\ 0, & C_A \text{ and } C_B \text{ do not have the same topic words} \end{cases}$$

Assuming there are N co-citation instances in a particular co-citation level, M of N co-citation instances are similar. The average similarity of the co-citation instances in this level is M/N, which means that if two references co-cited in this co-citation level, the similar probability of the co-cited references is M/N. The probability M/N would be treated as the weight of this co-citation level.

Compare the weighted co-citation analysis with traditional co-citation analysis

Co-citation clustering method was commonly used in co-citation analysis. The co-citation clusters were always used to reveal the research fronts of the citing papers. After weighting the co-citation strength at different co-citation levels, the co-citation clustering results might have some changes. These changes may also affect the identification of the research fronts.

Hierarchical clustering method was employed to cluster the co-citation papers. Co-cited papers with co-citation frequency 10 or more will be chosen for this experiment. The similarity of the co-cited papers was calculated with the following formula.

$$Similarity(D_A, D_B) = \frac{Frequency_{AB}}{\sqrt{Frequency_A \times Frequency_B}}$$

For the weighted co-citation clustering, the strength of the co-cited papers was calculated as follows.

$$Strength(D_A, D_B) = \sum_{i=1}^{4} (Weight(P_i) \times Frequency_{AB}(P_i))$$

Pi was the i level co-citation. (i=1: sentence level; i=2: paragraph level; i=3: section level; i=4: article level).
The similarity of the co-cited papers was calculated as follows.

$$Similarity(D_A, D_B) = \frac{Strength(D_A, D_B)}{\sqrt{Frequency_A \times Frequency_B}}$$

The clusters of the traditional co-citation will have some changes after weighting the co-citation strength. Some of the cited papers might move from one cluster to another. And then the citing papers of the clusters will change subsequently. The changes of the citing papers will be traced to reflect the influences of the moved paper. Figure 1 shows the changes of citing papers (S1) of the cluster (C1) when a cited paper (R1) joins to the cluster. When a new cited paper R1 joins to C1, papers which were citing both R1 and the papers in cluster C1 will compose a new set S2. Set S3 was the intersection of S1 and S2. The influence of R1 to C1was that topics of citing papers in S3 were strengthened. In other words, the size of S3 could reflect the effect of R1 to C1.

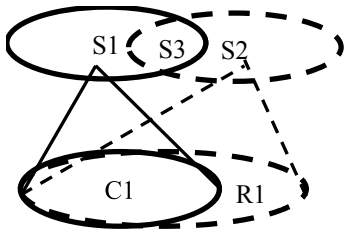


Figure 1. The influence of cited paper R1 to cluster C1

Results

The weight of different co-citation levels

Co-citations with frequency 10 or more in three BMC journals were extracted. The number of co-citation instances in each co-citation level was shown in table 1. The related rates of four co-citation levels were 1, 0.77, 0.64 and 0.56.

Table 1. The co-citation weight of four co-citation levels

	Sentence-level	Paragraph-level	Section-level	Article-level
Co-citation instances	**2131**	**1146**	**1150**	**2359**
Related co-citation instances	**2131**	**884**	**733**	**1321**
Related rates	**1**	**0.77**	**0.64**	**0.56**

The related rates will be treated as the weight of the co-citation levels. The weights were a little different from the weights in the papers of Gipp (Gipp, 2009b) and Callahan (Callahan, 2010). Gipp and Callahan gave a very small weight to the co-citations occurred in article level, such as 1/4. But in our research, the article level weight reaches 0.56.

Comparison of the weighted co-citation clustering and traditional co-citation clustering

The co-cited papers with the co-citation frequency 10 or more will still be used. The results of the hierarchical clustering showed two obvious changes between traditional co-citation clustering and weighted co-citation clustering. One is the paper "VON MERING C, 2002, NATURE, V417, P399" changes from one cluster to another after weighting the co-citations. The influence of this paper on the earlier cluster was 62.1%. And the influence on the later cluster was 86% which was better than it performed in earlier cluster. Another is the paper "BOECKMANN B, 2003, NUCLEIC ACIDS RES, V31, P365". The influence of this paper on the earlier cluster was 16.3% which was very low. And the influence on the later cluster was 22.2%. These results indicate that the weighted co-citation clustering performs better than the traditional co-citation clustering on identifying the topics of the citing papers.

Conclusion

We studied the similarity of the citation contexts of the co-cited papers in each co-citation level. The results were used as the co-citation weight of the co-citation level. The co-citation weights in 3 journals were 1, 0.77, 0.64, and 0.56. These weights were used to improve the co-citation clustering results. The results showed that the improved clusters were better in identifying the topic of the citing papers than traditional clusters.

Acknowledgments

This research is supported by National Natural Science Foundation of China (grant number 71003011) and Fundamental Research Funds for the Central Universities of China (Grant 852010).

References

Small, H. (1973). Co-citation in the scientific literature: A new measure of the relationship between two documents. *Journal of the American society for information science and technology*. 24,265-269.

Gipp, B & Beel, J. (2009). Identifying related documents for research paper recommender by CPA and COA. *Proceedings of international conference on education and information technology*(pp. 636-639). Berkeley: International Association of Engineers

Callahan, A., Hockema, S. & Eysenbach, G.(2010). Contextual Cocitation: Augmenting Cocitation Analysis and its Applications. *Journal of the American Society for Information Science and Technology*. 61,1130-1143

Elkiss, A., Shen, S., Fader, A., Erkan, G., States, D.& Radev, D. (2008) . Blind men and elephants: What do citation summaries tell us about a research article? *Journal of the American Society for Information Science and Technology*. 59, 51-62

Liu, S., Chen, C.(2012). The proximity of co-citation. Scientometrics. 91,495-511

WHAT MEANS, IN NUMBERS, A GOLD STANDARD BIOCHEMISTRY DEPARTMENT TO NATIONAL AGENCIES OF RESEARCH FOMENTATION IN BRAZIL?

Suelen Baggio; Ben Hur M. Mussulini; Diogo Losch de Oliveira; Diogo O. Souza; Luciana Calabró.

suzy-gfbp@hotmail.com; ben_hurmussulini@yahoo.com.br; diogolosch@gmail.com; diogo@ufrgs.br; luciana.calabro.berti@gmail.com
Universidade Federal do Rio Grande do Sul, Departamento de Bioquímica, Rua Ramiro Barcelos, 2600 – anexo, Porto Alegre, RS (Brazil)

Introduction

The postgraduation in Brazil has experienced remarkable growth. It was formally established in the mid-60s (Velloso, 2004), reaching the 70s with 500 postgraduate MSc program and 200 PhD program, according to CAPES (Brazilian funding agencie - Coordenação de Aperfeiçoamento de Pessoal de Nível Superior). Constantly growing, arrive at the year of 2010 with 2700 MSc and 1600 PhD programs. In the last global evaluation conducted by CAPES, in 2010, it was observed that in a scale of 1 to 7, with 7 considered the international standard performance, the average obtained by postgraduate programs (PGP) in Brazil was between 3 and 4, considered regular.

In the period from 1999 to 2003, Brazil was responsible for 1.5% of world production (going from 23 ° to 17 ° position), with the best ranking of Latin America (Berti et al., 2010). Among so many areas of scientific knowledge, Biochemistry has its highlight, presenting significant results that can be translated by the data presented by the Biochemistry PGP at UFRGS, with grade 7 (note 7 since 2001). To achieve 7, some points are needed, such as: good infrastructure, training human resources and scientific production. By analyzing these factors, the Biochemistry PGP presents annual increases in the number of scientific articles published and formation of masters and doctors.

In order to increase the international visibility of the UFRGS, as of the Biochemistry Department, and to better elucidate in numbers what means a note 7 by CAPES in our country, it was selected the six most productive researchers of this department (scientific articles and formation of Human Resources), and their scientific profiles were submitted to classical scientometrics analysis.

Methods

To achieve such goal, it was used the databases Scopus and Lattes (National Data Base), extracting the scientific profile of six researchers from the UFRGS Biochemistry Department (limit data analysis was 12/31/2012). The researchers were selected by number of publications and formation of human resources. The scientometric analyses applied in this study were: number of published articles; formation of PhD students; total citation; total self-citation; *h* index. In a last point, it was

designed a table summarizing the journals in which. often these researchers publish their results. The graphs were obtained by GraphPad Prism 5.

Results

As can be seen in Figure 1, the number of articles published by the researchers goes from 345 (teacher 1) to 144 (teacher 6). Concerning the formation of PhDs, Figure 2, teacher 1 formed 31 students, while the others formed from 17 to 11 PhD students.

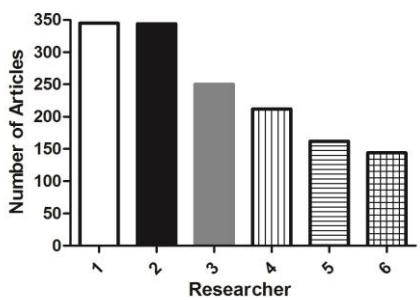


Figure 1. Number of scientific articles

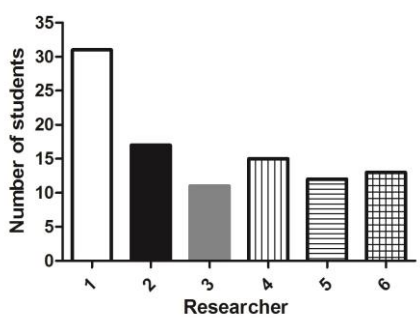


Figure 2. Number of students

Figure 3 shows the total number of citations, ranging from 5810 to 2268, and of self-citations, with values from 1525 to 288. The *h* index goes from 33 to 20.

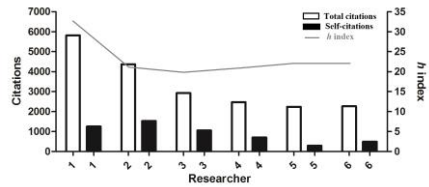


Figure 3. Number of total citations (white bars), self-citations (black bars), and *h* index of each researcher (grey line).

Among the journals in which the researchers mainly published their results, the higher impact factor is 3.577 (Journal of Inherited Metabolic Disease) and the smallest is 1.129 (International Journal Of Developmental Neuroscience) Table 1.

Table 1. Journals and correspondent impact factor (I.F), in which the researchers mainly published their results.

Journal	I. F.
Neurochemical Reserach	2,24
Brain Research	2,728
Neuroreport	1,656
Neurochemistry Internacional	2,857
International Journal Of Developmental Neuroscience	1,129
Metabolic Brain Disease	2,198
Brazilian Journal of Medical and Biological Research	2,418
Journal of Inherited Metabolic Disease	3,577
Molecular and Cellular Biochemistry	2,057
Free Radical Research	2,878
Journal of Medicinal Food	1,408
Toxicology in vitro	2,775
Behavioral and Neural Biology	1,984
Progress in Neuro Psychopharmacology and Biological Psychiatry	3,247
Clinical Biochemistry	2,076
Nueroscience Letters	2,105

Discussion

The UFRGS Biochemistry Department evolved over its 40 years. Among the factors to this development are the

arrival of international consolidated researchers, foundation of a postgraduate program, and fomentation to research from scientific agencies (Gomes et al. 2011). Fomentation agencies are increasingly using scientometric tools invest their limited resources. So, it is primordial to know what a scientific institution means in numbers. Such numerical evaluation help to better understand where a random institution should focus to develop, upgrading their visibility and therefore attracting more investments. In this context, this study elucidates what CAPES and other Brazilian institutions consider to perform investments.

In numbers, a gold standard biochemistry department in Brazil could mean per researcher: formation of 10 PhD students or higher; at least 150 scientific articles published; at least 2200 total citation; self-citation between 10-25% of total citation; *h* index 20 or higher; and mean impact factor 2.5. These numbers are based in our analysis of the six more productive researchers in the department. It is important to say that a more profound comparison among all Biochemistry Departments with note 7 by CAPES would be better, however the studies involving scientometric analysis of such institutions are scarce, making such comparison of difficult access.

Bias of such analysis could be negative citation, high degree of self-citation, great number of articles in low impact factor journals and the dilution of unique discoveries in high impact journals; nevertheless, analyzing each of this point are very difficult (Hirsch, 2005).

We hope that in light of these results, futures comparison among our department and international institutions could be performed to improve our scientific level, as well as to show for another national biochemistry department what is required to reach CAPES note 7.

Acknowledgments

This work was supported by the Brazilian funding agencies Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) and Instituto Nacional de Ciência e Tecnologia em Excitoxicidade e Neuroproteção (INCTEN).

References

Velloso, Jacques. (2004). Mestres e doutores no país: destinos profissionais e políticas de pós-graduação. *Cadernos de Pesquisa*, v. 34, n. 123, p. 583-611.

Luciana C. Berti, Diogo L. Oliveira, Diogo O. Souza, Susana T, Wofchuk. (2010). Produção científica e formação de recursos humanos na área de Bioquímica em Instituições federais do RS: fomento estadual. *Química Nova*, Vol. 33, No. 3, 765-771.

Urubatã E. Gomes, Diogo L. de Oliveira, Luciana C. Berti, Olavo Amaral, Diogo O. Souza, Susana T. Wofchuk. (2011). 37 years of scientific activity in a Biochemistry Departament in Brazil- patterns of growth and factors leading to incresead productivity. *Anais da Academia Brasileira de Ciências*, Vol. 83, nº 3, 1121-1130.

J. E. HIrsch. (2005). An index to quantify an individual´s scientific research output. *PNAS*, vol. 102, nº 46, 16569-16572.

lattes.cnpq.br

http://www.scopus.com/home.url

WHEN INNOVATION INDICATORS MEET SPIN-OFF COMPANIES: A BRIEF REVIEW AND IMPROVEMENT PROPOSAL

Zhiyu Hu

huzy@most.cn
Institute of Scientific and Technical Information of China, 15 Fuxing Road, 100038, Beijing (China)

Introduction

Spin-off companies play a key role in innovation system, benefited from strong links with academia and industry (Locketta *et al*, 2005). For research and management of innovation, it is of high priority to measure spin-offs with a set of practicable and informational indicators. However indicators for spin-offs in most research and statistics are insufficient to provide a panorama for investigators, policy makers and investors.

Up to date, most of the studies on spin-offs are macroscopic qualitative analysis. For example, Elpida *et al* (2010) build the conceptual framework of spin-off chain. Clarysse *et al* (2011) perform a comparative study on characteristics of corporate spin-offs and university ones. Furthermore, Finne *et al* (2011) report a composite indicator for general knowledge transfer.

Practically, spin-off companies have been only generally described in official guideline documents as part of the technology developing small and medium enterprises (SMEs), notably, *Oslo Manual* (EC, Eurostat, 2005). In the latest *EU Innovation Union Scoreboard 2011*, there are two indicators about them, SMEs innovating in-house (% of SMEs) and innovative SMEs collaborating with others (% of SMEs). However, both macro indicators are not specifically designed for spin-offs.

The Challenges faced by Innovation Metrics

Although the pilot qualitative studies make great progress of applying innovation indicators to spin-offs, there remains substantial mismatching problems which is attributable to spin-off's unique characteristics.

For spin-off, the process of production generally equals R&D. Product cost can be estimated by R&D expenses. This simply negates the basis of traditional indicator of R&D intensity and emphasizes the theoretical demand for a modified indicator system for spin-offs. Several aspects are to be discussed in details.

In most case, the purpose to establish a spin-off is commonly to improve the maturity of technology. Unlike traditional industry, there is no tangible product, nor even new patent in some case, from spin-off companies. Therefore, traditional output indicators based on product data, and integral part of innovation index system largely, may not provide an accurate readout for spin-offs.

Furthermore, in some of current indicator systems, 10 employees is set as the bottom line of firm's minimum size (Arundel, 2009), which is not applicable to a large part of spin-offs and results a

consequent loss of valuable data and a distorted macro-level picture.

Product, including knowledge, can be tangible or intangible. For spin-offs, intangible one is more routine and valuable. In this regard, many earlier censuses exhibit a tendency to apply indicators to mostly tangible capital, which prevents their direct application to spin-offs. For instance, total factor productivity (TFP), a popular indicator on macro-level, is apparently not optimal for spin-offs.

In summary, in spite of many pre-existing indicator systems, few of them are specifically designed for spin-offs. Given the unique properties of spin-off productivity, a modification of current innovation indicators is required and, therefore, proposed in the present study.

Improvement proposal for spin-off Indicators

Indicators for capabilities and resources

Input structure of spin-off can be really complicate. It's beneficial to evaluate capabilities or resources, rather than inputs.

Investment attracted can be a standard skill meter, as an indicator for capital. This information is probably available from survey of venture capitals or from associative organisations, such as business angel networks.

Another important indicator is management skills, which covers entrepreneurship, leadership and business skill of the core members. Lack of experienced managers is one of the most common reasons of spin-offs' failure. The *Canadian Survey on the Commercialisation of Innovation 2007* set a good example of application of management skill indicator.

Indicators for environment and linkages

As nonlinear interaction mode of innovation is generally accepted, indicators for linkage should be put in the heart of metrics for spin-offs.

In most research and practical works, there are four common indicators, Co-publications, Co-patenting, Licenses, and finally Contract R&D, even in case of relationship among companies. They are important but insufficient in the absence of other connections, such as business consulting, training, etc. Linkage indicators covering commercial linkages are necessary supplements to technical touch.

In addition, traditional indicators don't measure connections with business angles, or start-up capital. Angle investment information is likely available in trade publications or newsletters of associations.

Indicators for growth and potential

Most traditional surveys obtain outputs data for statistics. Recently, methodology switches from static to dynamic. It is necessary and valuable to move indicators from outputs (static) to growth and potential (dynamic) of a spin-off.

Empirical studies about firm growth should be highly appreciated, such as Geroski (2003) using a sample of 147 UK firms observed continually for more than 30 years. Zhao (2011) classify growth paths and factors affecting.

A slightly revised innovation survey metrics should be used to identify fast-growing spin-off or "gazelles". Speeds of employment changing or virtual market value growth will be meaningful indicators for evaluations.

Summary of all indicators

There is a summary of all indicators in Figure 1. All elements in three dimensions form a sturdy structural model. In each dimensions, several traditional innovation indicators integrate with proposed modified innovation indicators. This model may be applying to comprehensive quantitative evaluation for future research.

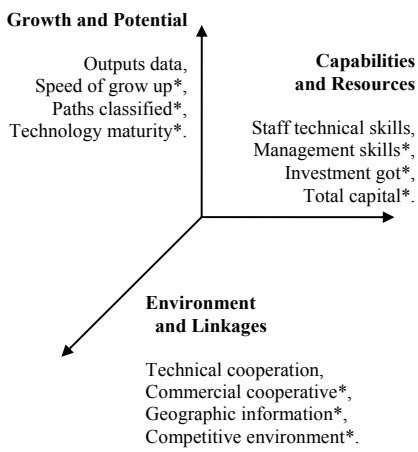


Figure 1. Model of Modified Innovation Indicators for Spin-offs. Proposed modified indicators marked with"*".

Potential applications

I propose that this modified set of innovation indicators for spin-off companies warrants further validation and modification in empirical research.

The application of modified innovation indicators will potentially increase the resolution of national or regional innovation surveys and research.

Moreover, modified innovation indicators can be applied to evaluate a given spin-off company so as to provide a more comprehensive spectrum of information to potential investors and investigators.

Acknowledgments

This work is supported by the National Natural Science Foundation of China.

References

Arundel, A., Lorenz, E., Lundvall, B. Å., & Valeyre, A. (2007). How Europe's economies learn: a comparison of work organization and innovation mode for the EU-15. *Industrial and Corporate Change*, 16(6), 1175-1210.

Arundel, A., O'Brien, K. R. (2009). *Innovation metrics for Australia.*

Clarysse, B., Wright, M. and Van de Velde, E. (2011). Entrepreneurial Origin, Technological Knowledge, and the Growth of Spin-Off Companies. *Journal of Management Studies*, 48(6), 1420–1442.

Elpida, S., Galanakis, K., Bakouros, I., & Platias, S. (2010). The Spin-off Chain. *J. Technol. Manag. Innov.*, 5(3), 51-68.

European Commission, Eurostat, (2005). Oslo Manual: Guidelines for Collecting and Interpreting Innovation Data, 3rd Edition. Paris: OECD Publishing

European Union, (2012). *Innovation Union Scoreboard 2011*. Brussel: EU Publishing

Finne, H., Day, A. & Piccaluga, A, etc. (2011). A Composite Indicator for Knowledge Transfer, Report from the European Commission's Expert Group on Knowledge Transfer Indicators. Trondheim

Geroski, P. A., Lazarova, S., Urga, G. etc. (2003), Are differences in firm size transitory or permanent? *J. Appl. Econ*., 18: 47–59.

Locketta, A., Siegelb, D. & Wrightc, M., etc. (2005). The creation of spin-off firms at public research institutions: Managerial and policy implications. *Research Policy*, 34(7), 987-983.

Zhao, C., Wang, J., Zhou, Q., (2011), Research oil the enterprise growth path from the perspective of intellectual capital: A case study of high·tech SMEs. *Business Economics and Administration*, (12):61-69

WHERE NATURAL SCIENCES (PHYSICS) MADE IN THE WORLD AND IN RUSSIA: 3-DECADES DYNAMICS

Michael Romanovsky

slon@kapella.gpi.ru
A.M.Prokhorov General Physics Institute, Vavilov str., 38, 119991 Moscow (Russia)

Introduction

There the principal question what form of research organization is more successful with respect to scientometric indicators like the number of publication and citation per one researcher, etc. It is clear that the general cross-consideration among all countries involved into research process contains several principal difficulties. The first one is the language factor. Non-English speaking scientific journals as well as non-English speaking countries are in worth conditions with respect to citation. Attempts to equalize these conditions have led (in continental Europe) to the disappearing of scientific journal on national languages like Nuovo Cimento, etc. Nevertheless, the English language of the scientific journal does not provide the rise of impact-factor of the journal immediately. Correspondingly, the country that starts to publish scientific papers in English does not esquire any preferences automatically, and newly-Anglicized scientific journals do not esquire soon the addition impact-factor. Above that, publication traditions are varied from country to country: somewhere, it is convenient to publish scientific results soon, somewhere – not. As a result, leading impact-factor journals are publishing in the USA, and UK. Scientists form these countries as well as other English-speaking countries have some advantages in comparison with scientists from other countries.

The financing of scientific groups is another powerful cause of the difference in quantity and "quality" of scientific papers. This factor may be strongly changed from country to country and not so strongly from one scientific group to another one inside one country. Thus there is big difference in direct comparisons of scientists and scientific groups in different countries. On the other hand, this difference goes away if scientists and scientific groups of one country are estimated using above scientometric indicators. Moreover, they can be estimated directly using publication in leading foreign scientific journals. For non-English speaking countries it can be leading journal like Nature, Science, etc. We will analyze what "part of science" is made in universities, and in other research organizations. Specifically for Russia, we will analyze three types of research organizations: Russian Academy of Sciences, universities, and others.

Method of analysis [1]

We will analyse the number of publications in two scientific journals: Nature (only the main issue not special ones), and Physical Review Letters (PRL). The choice of these journals was due to the fact that they exist in present view during all period of analysis. Three decades: 1981-1990, 1991-2000, and

2001-2010 have been chosen. We use data from ISI Web of Knowledge. Name of journals (Nature or Physical Review Letters), country name, and the decade (for example, 1981-1990) were indicated there. Scientific articles were selected only. As a result, the table with names of research organizations, number of publications of these organizations, and the part of these organizations was generated.

This table contained not only the country under analysis since ISI Web of Knowledge accounts all countries of co-authors. There was the crucial assumption to consider research organizations from the considered country and discard all others. The error introduced by this assumption may arise due to various distributions of co-authors number for different research organizations. We suppose that this error did not significant. Note that the number of co-authors may vary with respect to the type of research organization. Indeed, such organizations like Max-Plank-Society in Germany, National laboratories in the USA, Russian Academy of Sciences are more all-sufficient (in facilities, materials, etc.) in comparison with universities. Therefore, scientists from such research organizations do not need wide co-authoring of their scientific articles. The last wide co-authoring leads to some overstating of articles from universities in final results.

Results

The analysis of origin of articles from Germany and the USA was done for journals Nature and PRL provided the answer "What part of (fundamental) physics is made in universities and other research organization?" for decades 1981-1990, 1991-2000, and 2001-2010. Figure 1 presents the parts of articles from the USA in Nature and PRL from Universities and other organization (4 triples), the dark left column in each triple represents 1981-1990 period, more hell central column – 1991-2000, hell right column – 2001-2010:

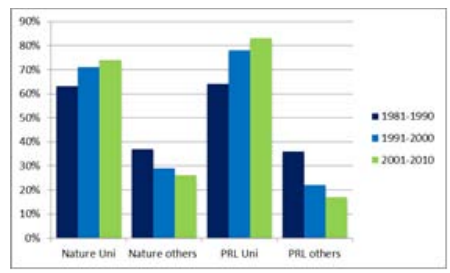


Figure 1. Parts of articles in Nature and PRL from the USA Universities and other organizations in Nature and PRL (4 triples).

It is seen the drop of publications part of other research organizations in time. It may connect with the noted fact of more wide co-authoring of papers from universities since universities are not so all-sufficient in fundamental investigations. The research facility became more complicated during last 30 years and now is too expensive for the middle USA university. The drop of articles in PRL of part of non-university research organizations in period of 30 years is stronger than in Nature since physics is the most expensive science.

The same analysis was done for German research organizations, results were slightly different from USA ones. For all natural sciences published articles in Nature, the university part became larger than the part of other organizations even in this century. Other research organizations in Germany are included in societies like Max-Plank-Society, Leibniz Association, etc. mostly. From the other hand, large number of publications from medical centres like "Charite" in Berlin belonged

to universities and accounted in the university part (see Figure 2):

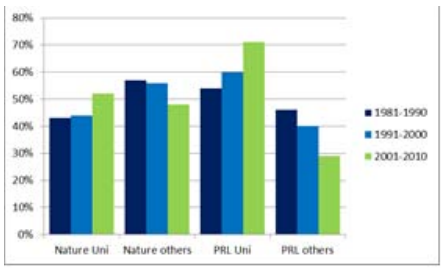


Figure 2. Parts of articles in Nature and PRL from Universities and other research organizations of Germany (the structure is the same that in Figure 1).

Thus the answer on the question "Where is the fundamental science made in Germany?" is not so clear that for the USA. At the same time, German publications in PRL clearly drop in time as it is in the USA. This drop of physical publications from non-university research organizations in Germany may be connected with the above speculation of widening of co-authoring in German universities like in the USA.

The situation with fundamental investigations in Russia is differed from the same in other countries due to the existence of the strong scientific centre: Russian Academy of Sciences (RAS). Thus the analysis was done for 3 types of organizations: RAS, universities, and others (dark, medium, and hell columns in each triple of Figure 3). For physics, the last type of organizations is represented by National Centres like Kurchatov Institute, National Research Centre in Sarov, etc. The left column in each triple of Figure 3 represents 1981-1990 period, the middle column – 1991-2000, the right column – 2001-2010:

The university part and part of other organizations are small. It means that natural sciences are made in Russia in RAS mostly. For physics, Figure 3 says that physics is produced in other research organizations of Russia mostly in a new age in comparison especially with last decade of the USSR: it is connected with the closed character of conducted work in other research organizations.

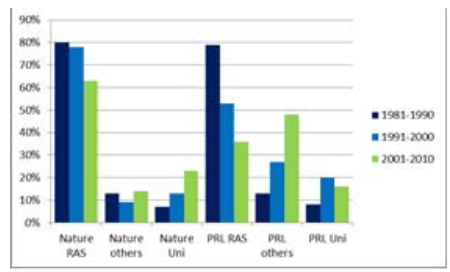


Figure 3. Parts of articles in Nature and PRL of RAS, other organizations, and Universities in Russia.

Conclusion

The most part of fundamental investigation in natural sciences in the USA is fulfilled in universities, this part is rising during the period of observation (1980-2010). The same rise in physics is more evident. The same analysis fulfilled for Germany demonstrates that it is hard to claim that universities are leading research organizations for natural sciences in Germany, but their part is rising. Natural sciences are investigated in Russia by RAS and non-university research organizations.

Acknowledgments

The work is supported by RFBR grant 13-07-00672-a.

References

Romanovsky, M.Yu. (2010) Publication activity of natural-science research organizations in Russia and abroad. *Herald of the Russian Academy of Sciences*, 80, 475-479.

AUTHOR INDEX

Abbasi, Alireza 328
Abdiazar, Shahram 1995
Abdulhayoglu, Mehmet Ali 1151
Abercrombie, Robert K. 1854
Abramo, Giovanni 1536
Acosta, Manuel 36
Adams, Jonathan 316
Aguillo, Isidro F. 1966, 2159
Ajiferuke, Isola 755
Aksnes, Dag W. 654
Albarrán, Pedro 536
Aleixandre-Benavent, Rafael 1932
Amaral, Roniberto M. 1877
Amez, Lucy 1891
Andersen, Jens Peter 215
Antonio-García, M. Teresa 2149
Aparicio, Javier 2044
Arali, Uma B. 2117
Archambault, Éric 1665
Arencibia-Jorge, Ricardo 2113
Asadi, Hamideh 2017
Åström, Fredrik 677
Atanassova, Iana 591
Azagra-Caro, Joaquín M. 36
Babko-Malaya, Olga 896
Badrloo, Alireza 2017
Baggio, Suelen 2193
Barberio, Vitaliano 426
Barbosa, Nilda Vargas 1884
Bar-Ilan, Judit 468, 604
Barirani, Ahmad 1944
Barrios, Maite 811, 966, 1922, 2156
Barth, Andreas 493
Basner, Jodi 2066
Basu, Aparna 1954
Batista, Pablo Diniz 796
Bauer, Hans P.W. 2099
Benoit, Cyril 1947
Bertin, Marc 591
Besagni, Dominique 2048
Bharathi, D. Gnana 58
Bi, Fei 1069
Bidanda, Bopaya 1225
Bleier, Arnim 229, 2171
Bollen, Johan 3
Bonaccorsi, Andrea 1817
Bongioanni, Irene 1900
Bordons, María 167, 2044, 2162
Borges, Elinielle Pinto 796
Börner, Katy 1342, 1587
Bornmann, Lutz 493, 769
Bouabid, Hamid 885
Bouyssou, Denis 2024
Boyack, Kevin W. 361, 928, 1726
Bozeman, Barry 1613
Bucheli, Víctor 1225
Buckley, Kevan 1253
Büsel, Katharina 175
Cabezas-Clavijo, Álvaro 96, 1237
Cabral, José A. S. 2054
Cabrini Grácio, Maria Cláudia 1908, 2069
Căbuz, Alexandru I. 2086
Cainarca, Gian Carlo 2004
Calabró, Luciana 1884, 2129, 2193
Calderón, Juan Pablo 1225
Cambo, Scott Allen 1711
Cárdenas-Osorio, Jenny 1928
Carley, Stephen J. 1188
Cassi, Lorenzo 1270
Castellano, Claudio 769, 1431
Castellano-Gómez, Miguel 1932
Chang, Chia-Lin 1871
Chavarro, Diego 1053
Chen, Bikun 742
Chen, Chaomei 847, 1037, 1114, 1726
Chen, Dar-Zen 941, 1379
Chen, Ssu-Han 941
Chen, Yunwei 1135
Cheng, Qikai 1307, 2178

Cherraj, Mohammed 885
Chi, Pei-Shan 612
Chinchilla-Rodríguez, Zaida 2061, 2113
Chittó Stumpf, Ida Regina 1935
Chmelařová, Zdeňka 1874
Chung, Kon Shing Kenneth 328
Cointet, Jean-Philippe 285
Colebunders, Robert 2072
Corera-Alvarez, Elena 2113
Coronado, Daniel 36
Corrigan, James 1485
Cosculluela, Antonio 2156
Costas, Rodrigo 84, 876, 1401, 1587
Crabtree, Dennis R. 2092
Cronin, Blaise 1321, 1640
D'Angelo, Ciriaco Andrea 1536
da Costa Santos, Maria José Veloso 2174
da Silveira Guedes, Vânia Lisboa 2174
Dafang, Tian 1912
Dalimi, Mohamed 885
Damasio, Edilson 1925
Daraio, Cinzia 1817, 1900, 2004
das Neves Machado, Raymundo 1759
de Faria, Leandro I. L. 1877, 1950
de Fátima Sousa de Oliveira Barbosa, Maria 2174
De Filippo, Daniela 1868, 2095
de Magalhães Mollica, Maria Cecília 2174
de Moya-Anegón, Félix .. 2061, 2113
de Nooy, Wouter 769
de Oliveira, Diogo Losch 2193
de Souza Vanz, Samile Andréa 1935
de Souza, Diogo O. G. 2129
Degelsegger, Alexander 175, 177, 183
Dehdarirad, Tahereh 1922
Dekleva Smrekar, Doris 1976
Demarest, Bradford 2027
den Besten, Matthijs 484
Deritei, Dávid 2086
Derrick, Gemma Elisabeth 136
Díaz-Faes, Adrián A. 2162
Didegah, Fereshteh 1830, 1995
Digiampietri, Luciano A. 447
DiJoseph, Leo 1485
Ding, Jielan 1177
Ding, Kun 1114, 2020, 2189
Ding, Ying 264, 1030, 1106
Diwakar, Sandhya 1963, 2089
do Amaral, Roniberto M. 1363, 1950
Doleželová, Jana 1874
Dong, Huei-Ru 1379
Dong, Ke 339
Dorta-González, María Isabel 1847, 2146
Dorta-González, Pablo ... 1847, 2146
Du, Qing 1528
Duanyang, Xu 1938
Egghe, Leo 1159
Engels, Tim C. E. . 1170, 1861, 1894
Ercsey-Ravasz, Mária 2086
Escribano, Alvaro 978
Fan, Chun-liang 551
Fanelli, Daniele 2080
Fang, Shu 1135
Faria, Leandro I. L. 1363
Fayazi, Maryam 2031
Finstad, Samantha 1485
Florian, Răzvan V. 2086
Fornieles, Albert 2156
Franceschini, Fiorenzo 300
François, Claire 2048
Fritsche, Frank 1989
Gallié, Emilie-Pauline 1270
Galvis-Restrepo, Marcela 1928
Gani, Srishail 2117
Ganji, Mahsa 2017
Garcia Romero, Antonio 978
García-Zorita, J. Carlos 418, 2095, 2126
Garzón-García, Belén 2149

Gaughan, Monica 1613
Getz, Daphne 1970
Gholami, Nima 2017
Giménez-Toledo, Elea 1861
Gingras, Yves 591
Glänzel, Wolfgang 109, 237, 1864, 2080
Gomes, José A. N. F. 2054
Gomez-Benito, Juana 966
Gómez-Nuñez, Antonio J. 2061
Gómez-Sánchez, Alicia F. 1973
Gomila, Jose M. Vicente 861
González, Fabio 1225
González-Albo, Borja 2044
González-Teruel, Aurora 2156
Goodarzi, Samira 2017
Gornstein, Luba 1019
Gorraiz, Juan 519, 626, 1237
Gorry, Philippe 1947
Graffner, Mikael 677
Greenspan, Emily J. 1485
Gregolin, Jose A. R. 1363, 1877, 1950
Guerrero-Bote, Vicente P. 1469
Guilera, Georgina 966
Gumpenberger, Christian 519, 626, 1237
Guns, Raf 353, 819, 1409
Guo, Ying 1278, 2083
Guo, Yu 1069
Gupta, Mona 1963, 2057
Gurney, Karen 316
Gurney, Thomas 1090
Gutierres Castanha, Renata Cristina 1908
Haddow, Gaby 1210
Hammarfelt, Björn 720
Han, Shuguang 1307
Han, Yi 377
Hatami, Mahdieh 2017
Haustein, Stefanie 468
Havemann, Frank 1881
Hayashi, Kazuhiro 1905
He, Jiangen 2178
Heck, Tamara 1392
Hedenfalk, Ingrid 677
Heeffer, Sarah 1864
Hefetz, Amir 1970
Henriksen, Dorte 152
Herrera, Francisco 1550
Ho, Yuen-Ping 635, 1622
Holmberg, Kim 567
Holste, Dirk 2048
Hong, Lv 2182
Hook, Daniel 316
Hopkins, Michael 251
Hörlesberger, Marianne .. 1738, 2048
Horlings, Edwin 1090
Hou, Haiyan 1792, 1941
Hou, Jianhua 1941
Hsu, Elizabeth 1485
Hu, Qing-Hua 272
Hu, Zewen 2165
Hu, Zhigang 847, 1941
Hu, Zhiyu 2196
Hua, Weina 2102
Huang, Mu-Hsuan 941, 1379
Hui, Xia 377
Hunter, Daniel 896
Ikeuchi, Atsushi 728
Ingwersen, Peter 418, 1003, 2126
Isabel-Gómez, Rebeca 1973
Ishtiaque Ahmed, Syed 1711
Itsumura, Hiroshi 1772
Iwami, Shino 507
Jack, Kris 626
Járai-Szabó, Ferenc 2086
Jarneving, Bo 955
Jensen, Unni 2066
Jiang, Chunlin 1941
Jiménez-Contreras, Evaristo 96, 1237
Jo, Karen 2066
John, Marcus 1989
Jonkers, Koen 136
Julian, Keith 1357
Jung, Hyosook 2152
Junpeng, Yuan 1887

JunPing, Qiu 2182
Juznic, Primoz 1976
Kajikawa, Yuya 507, 2034, 2037
Katranidis, Stelios..... 1334
Kay, Luciano 1202
Keidar, Yifat..... 2040
Kenekayoro, Patrick 1253
Khadka, Alla G. 690
Kitt, Sharon..... 1746
Klavans, Richard 361, 928, 1726
Kong, Xiangnan..... 1030
Koukliati, Olga 2120
Kousha, Kayvan 705, 1966, 2017
Kraker, Peter..... 626
Krampen, Günter 2099
Kreuchauff, Florian 1291
Kumar Srivastava, Vijai 2075
Lagoze, Carl 1711
Lahatte, Agenor 1270
Lampert, Dietmar 175
Larivière, Vincent.....
..... 591, 1321, 1640, 1897
Larsen, Birger.....
..... 418, 1003, 1881, 2126
Lascurain-Sánchez, Maria-Luisa
..... 2126
Laurens, Patricia..... 1090
Lázár, Zsolt I. 2086
Leck, Eran..... 1970
Lee, Jerry S.H. 1485, 2066
Lee, Jongwook..... 2051
Lei, Shengwei 2178
Leino, Yrjö 1992
Leng, Fuhai..... 404
Lepori, Benedetto 426
Leta, Jacqueline 796, 1759
Levitt, Jonathan M. 1461
Lewison, Grant 1601
Leydesdorff, Loet .....
..... 251, 316, 769, 1037
Li, Xiao-xuan..... 551
Li, Xin 1857
Li, Yu..... 2102
Li, Yunrong 1431
Lietz, Haiko..... 1566
Light, Robert P. 1342
Lin, Yen-chun 1918
Lipitakis, Evangelia A. E. C..... 22
Liu, Qing 1177
Liu, Shengbo 1114, 2189
Liu, Wen-bin 551
Liu, Xiang 831
Liu, Yu 2020
Liu, Yuxian 819, 1696
Liu, Zeyuan 847
Lopez Illescas, Carmen 136
López-Cózar, Emilio Delgado.. 1550
López-Navarro, Irene 2149
Lu, Kun 755, 2178
Lu, Wei..... 1307, 2178
Lu, You-min..... 1918
Luan, Chunjuan 1792
Lyu, Peng-hui..... 831
Ma, Fei-cheng 831
Ma, Jianxia 1857
Ma, Mingguo..... 1857
Ma, Zheng 1069
Macaluso, Benoit..... 1321
Macedo, Thiago D. 1877
Magalhães, Jorge L 2185
Maisano, Domenico 300
Maissonneuve, Nicolas..... 484
Maiwald, Gunar..... 2008
Maleki, Ashraf..... 2017
Mañana-Rodríguez, Jorge 1960
Maraut, Stéphane..... 484
Marchant, Thierry 2024
Mardani, Amir Hossein 1995
Martinez, Catalina 484
Marugan, Sergio..... 2095
Marx, Werner 493
Mas-Bleda, Amalia 1966
Mastrogiacomo, Luca..... 300
Matoh, Robert 1976
Mauleón, Elba 167, 1868, 2004
Mayr, Philipp..... 1493
McAleer, Michael 1871
McCain, Katherine W. 185

Mehdizadeh-Maraghi, Razieh .. 2017
Mena-Chalco, Jesús P. 447
Mendez-Vasquez, Raul Isaac ... 2132
Merindol, Valérie 1270
Meyers, Adam 896
Michels, Carolin 2105
Midorikawa, Nobuyuki..... 1983
Mier, Zhang 2011
Mikulka, Thomas..... 1237
Milanez, Douglas H. .....
..... 1363, 1877, 1950
Milanez, Mateus G. 1950
Milojević, Staša 264, 1106, 1321
Mingers, John C..... 22
Minguillo, David 985
Miyairi, Nobuko 1905
Mohammadi, Ehsan..... 200
Mongeon, Philippe 1897
Montalt, Vicent..... 1932
Moore, Nicole M. 2066
Moreno, Luz 2044
Moreno-Torres, Jose Garcia 1550
Mori, Junichiro 507, 2034
Moshtagh, Shadi..... 2017
Moya-Anegón, Félix..... 1469
Mugnaini, Rogério..... 447, 1578
Muhonen, Reetta..... 1992
Munoz-Ecija, Teresa..... 2061
Murugan, M. Anand 1915
Mussulini, Ben Hur M. 2193
Nagahara, Larry A. 2066
Nakajima, Ritsuko 1983
Nakamura, Hiroko 2037
Ni, Chaoqun..... 1979
Nilbert, Mef..... 677
Noyons, Ed 1210, 1587
Olensky, Marlies..... 1850
Olinto, Gilda..... 796
Olivé Vázquez, Gerbert 2132
Ollé, Candela 811
Onofre Souza, Diogo 1884
Ortega, Lidia..... 811, 2156
Ossenblok, Truyken L.B..... 1894
Oxley, Les..... 1871
Ozel, Bulent..... 2124
Pan, Yuntao 1069
Panagiotidis, Theodore..... 1334
Papp, István 2086
Pardo, Daniel..... 1090
Park, Seongbin 2152
Patil, Chandrashekhar G..... 2117
Perianes-Rodriguez, Antonio 536
Peritz, Bluma C. 1019
Pero, Mickael 912
Peters, Isabella..... 468
Pinto de Miranda, Elaine Cristina
..... 1578
Polanco, Xavier 2109
Polley, David E. 1342
Ponomarev, Ilya 2066
Porter, Alan L. .....
.....861, 1188, 1278, 2083
Pouris, Anastassios..... 2014, 2120
Pouris, Androniki 2014
Priem, Jason 468
Puuska, Hanna-Mari..... 1992
Qiu, JunPing..... 339, 2001
Quist, Galena..... 2143
Quoniam, Luc..... 2185
Radicchi, Filippo 769, 1431
Rafols, Ismael..... 251, 1037, 1053
Raj, Aparna Govind 2075
Reijnhoudt, Linda..... 1587
Rey-Rocha, Jesús 2149
Riechert, Mathias 1566
Rigby, John 1357
Rimmert, Christine..... 1957
Rimmert, Edith..... 1957
Rivera-Torres, Sandra Carolina 1928
Robinson, Douglas K.R.. 1278, 2083
Robinson-García, Nicolás .....
..... 96, 1237, 1550
Roche, Ivana..... 2048
Roe, Philip..... 1601
Romanovsky, Michael..... 2200
Rongying, Zhao..... 77
Rons, Nadine..... 1998
Roos, Daniel Henrique ... 1884, 2129

Rotchild, Nava 2040
Rotolo, Daniele 251
Rousseau, Ronald 1409, 2072
Ruibin, Wei 1912
Ruiz-Castillo, Javier 536, 1431
Ruocco, Giancarlo 1900
Rybachuk, Victor 2143
Safonova, Maria 389
Sakata, Ichiro 507, 2037
Sandström, Ulf 664, 2140
Sangam, Shivappa L. 2117
Sanz-Casado, Elias . 418, 2095, 2126
Schaer, Philipp 1392
Scharnhorst, Andrea 1587
Schiebel, Edgar 1419, 2048
Schlicher, Bob G. 1854
Schlögl, Christian 519, 626
Schmitt, Marco 1986
Schneider, Jesper W. 152
Schnell, Joshua D. 1485, 2066
Schoen, Antoine 1090
Schoeneck, David J. 1188
Schui, Gabriel 2099
Schulz, Jan 1784
Schwechheimer, Holger 1957
Seeber, Marco 426
Seger, Yvette R. 1485
Sepehr-Ara, Parisa 2017
Serrano-López, Antonio Eleazar 418, 2126
Shan, Shi 1445, 2001
Shao, Liming 1938
Sheldon, Frederick T. 1854
Shema, Hadas 468, 604
Shengnan, Wu 77
Shi, DingHua 1445
Shirabe, Masashi 123
Simar, Léopold 1817
Simon, Johannes 175
Singh Kushwah, Arvind . 1963, 2057
Singh, Keshari K. 2089
Sirtes, Daniel 784
Sivertsen, Gunnar 654, 1861
Siyahi, Akram 2017
Small, Henry 928
Smith, Alastair G. 1806
Sokolov, Mikhail 389
Sorensen, Aaron A. 1726
Souza, Diogo O. 2193
Srivastav, Ajay Kumar 1915
Srivastava, Divya . 1963, 2057, 2075
Stark, Abigail R. 690
Steenrod, Johanna E. 690
Strotmann, Andreas 229, 1082, 2171
Struck, Alexander 2168
Suerdem, Ahmet 2124
Sugimoto, Cassidy R. 264, 1106, 1321, 1640, 1979, 2027
Sulima, Pawel 2066
Sumi, Róbert 2086
Sumikura, Koichi 1090
Suñén-Pinyol, Eduard 2132
Suominen, Arho 1506
Suya, Hu 2011
Suzuki, Shinji 2037
Suzuki, Takafumi 728
Takei, Chizuko 728, 1772
Tan, Xin 1528
Tang, Puay 1053
Tannuri de Oliveira, Ely Francina 2069
Tavakoli, Mohsen 2017
Teichert, Nina 1291
Teixeira da Rocha, João Batista 1884, 2129
Terliesner, Jens 468
Thelwall, Mike 200, 567, 604, 705, 985, 1253, 1321, 1461, 1830, 1966
Thijs, Bart 237, 1151, 1864
Thomas, Patrick 896
Tijssen, Robert J.W. 583
Toivanen, Hannes 1506
Tong, Ying 377
Torres-Salinas, Daniel 96, 1237, 1550
Tribó, Josep A. 978
Tsay, Ming-yueh 1918

Tschank, Juliet 175
Tsou, Andrew 264
Tsuji, Keita 728
Turbany, Jaume 2156
Valderrama-Zurian, Juan Carlos 1932
Valdivia, Juan Alejandro 1225
van den Besselaar, Peter 136, 664, 1090
van Eck, Nees Jan 455, 1649
van Leeuwen, Thed N. 66, 654
Vanoiee, Sheida 2017
Vargas-Quesada, Benjamín 2061
Velden, Theresa 1711
Verhagen, Marc 896
Verleysen, Frederik T. 1170
Vieira, Elizabeth S. 2054
Vila Domènech, Joan Salvador 2132
Villarroya, Anna 811, 1922
Waaijer, Cathelijn J.F. 7
Wagner, Isabella 175
Waltman, Ludo 455, 1649
Wang, Bo 1114, 2189
Wang, Juan 1528
Wang, Qi 2140
Wang, Xianwen 1792
Wang, Xiaoguang 1307
Wang, Xuemei 1857
Wang, Yanling 2136
Wei, Guo 2011
Wei, Wenjie 819
Wepner, Beatrix 1738
Wernisch, Ambros 1237
Williams, Duane E. 1485
Wilson, Paul 1830
Winnink, Jos J. 583
Wolfram, Dietmar 755
Wong, Chan-Yuan 635
Wong, Poh-Kam 635, 1622
Wouters, Paul 66, 455, 876
Wu, Yishan 2165
Xu, Kan 1114
Yamaguchi, Kiyohiro 2034
Yamashita, Yasuhiro 1681
Yan, ChunNing 2001
Yan, Erjia 1030, 1106
Yang, Guo-liang 551
Yang, Liying 551, 1177
Yang, Sojung 2152
Yang, Yang 1887
Yegros-Yegros, Alfredo 84, 1401
Yin, Jiahui 819
Yishan, Wu 1887, 1912
Yitzhaki, Moshe 2040
Yoshikane, Fuyuki 728, 1772
Yoshinaga, Daisuke 1681
Youtie, Jan 1613
Yu, Guang 272
Yu, Tian 272
Yuan, Junpeng 1938
Yuntao, Pan 1887
Zahedi, Zohreh 876
Zanotto, Sônia Regina 1935
Zarama, Roberto 1225
Zhai, Lihua 1069
Zhang , Zhiqiang 1857
Zhang, Lin 237
Zhang, Ling 1528
Zhang, Yanan 1528
Zhang, Yi 861
Zhang, YiFei 1445, 2001
Zhao, Dangzhi 1082
Zhao, Rongying 742
Zhou, Qiuju 404
Zhou, Xiao 861, 1278, 2083
Zitt, Michel 285
Zontanos, Costas 1334
Zuccala, Alesia 353
Züger, Maria-Elisabeth 1419

Partners

ISBN: 978-3-200-03135-7
ISSN: 2175-1935