Support Programs to Increase the Number of Scientific Publications Using Bibliometric Measures: The Turkish Case

Yaşar Tonta[1]

[1] *yasartonta@gmail.com*
Hacettepe University Department of Information Management, 06800 Beytepe, Ankara (Turkey)

"*Not everything that counts can be counted, and not everything that can be counted counts*." – William Bruce Cameron

Abstract
Bibliometric measures for scientific journals such as journal impact factor, cited half-life, and article influence score are readily available through commercial companies such as Thomson Reuters, among others. These metrics were originally developed to help librarians in collection building and are based on the citation rates of published papers. Yet, they are increasingly being used, albeit undeservedly, as proxies for peer review to assess the quality of individual papers; and research funding, hiring, academic promotion and publication support policies are developed accordingly. This paper reviews the use of such metrics by the Turkish Scientific and Technological Research Council (TUBITAK) in its Support Program of International Scholarly Publications and concentrates on the most recent policy changes. A sample of 228 journals was selected on the basis of stratified sampling method to study the impact of changing algorithms on the level of support that journals received in 2013 and 2014. Findings are discussed and some recommendations are offered to improve the existing algorithm.

Conference Topic
Country level studies

Introduction
Bibliometric measures such as journal impact factor (JIF) and cited-half life are based on citation rates of published papers in the literature and their aging. They were originally developed to help librarians in collection building and in making decisions as to how long the back issues of journals should be kept in stacks (San Francisco, 2012). Yet, such bibliometric measures are often used to assess the quality of individual papers, authors, and institutions. They are increasingly being used, albeit undeservedly, as proxies for peer review to assess the quality of individual papers; and research funding, hiring, academic promotion and publication support policies are developed accordingly. Algorithms used to rank authors, institutions or even countries are primarily based on such bibliometric measures as JIF and h index (Simons, 2008). This paper reviews the use of such metrics by the Turkish Scientific and Technological Research Council (TUBITAK) in its Support Program of International Scholarly Publications and concentrates on the most recent policy changes.

Literature Review
The drawbacks of citation-based metrics, especially JIF, for research assessment is well documented in the literature (e.g., Seglen, 1997; Guerrero, 2001; Simons, 2008; Browman & Stergiou, 2008; Lawrence, 2008; Todd & Ladle, 2008; Balarama, 2013; Kotur, 2013; Marks, Marsh, Schroer & Stevens, 2013; Marx & Bornmann, 2013; Casadevall & Fang, 2014; Jawaid, 2014). Convincing arguments supported by empirical data were brought forward as to why such measures should not be used to evaluate research (e.g., skewed citation distributions, different publication and citation practices in Science vs. Social sciences, and the manipulation of JIFs by editorial policies). Some researchers stressed the hidden dangers

of a "citation culture" (Todd & Ladle, 2008) while others drew attention to how measurement and "bean counting" harms science (Lawrence, 2008), as such metrics can easily be "gamed" (Marks et al., 2013). The title of the editorial of the special issue on "the use and misuse of bibliometric indices in evaluating scholarly performance" of the journal *Ethics in Science and Environmental Politics* says it all: "Factors and indices are one thing, deciding who is scholarly, why they are scholarly, and the relative value of their scholarship is something else entirely" (Browman & Stergiou, 2008).

The San Francisco Declaration on Research Assessment (DORA), signed by researchers, journal editors and publishers alike, strongly recommends not to use "journal-based metrics, such as Journal Impact Factors, as a surrogate measure of the quality of individual research articles, to assess an individual scientist's contributions or in hiring, promotion, or funding decisions" (San Francisco, 2012). "[M]ost experts agree that the JIF is a far from perfect measure of scientific impact" (Bollen, Van de Sompel, Hagberg & Chute, 2009). Even Thomson Reuters, the publisher of such metrics through its Journal Citation Reports (JCR), is against using JIF to measure the quality of scientific papers (Marx & Bornmann, 2013, pp. 62-63). Yet, its use as "a tool of research assessment has reached epidemic proportions worldwide, with countries like India, China and the countries of Southern Europe being among the hardest hit" (Balaram, 2013, p. 1268). Some declared war on the impact factor (Balaram, 2013) and advised that its use should be abolished (Hecht, Hecht & Sandberg, 1998). Nonetheless, it is believed that, despite its misuse and abuse, JIF "will retain its impact and won't fade away" (Jawaid, 2014).

Consequently, policies developed for hiring, academic promotion, research funding, and monetary support to scientific publications in different countries tend to rely increasingly on metrics based on citation rates of published papers. Turkey is no exception (Tonta, 2014). The Higher Education Council of Turkey (YÖK) and the Turkish Scientific and Technological Research Council (TUBITAK) have been using journal impact factors for almost two decades in their academic promotion policies and incentive programs to support scientific papers, respectively.

The use of bibliometric measures for research assessment in Turkey along with their suitability as criteria to evaluate research quality has recently been reviewed (Tonta, 2014). This paper examines the most recent algorithmic changes introduced in 2013 and 2014 to rank the journals in the Support Program of International Scholarly Publications (UBYT) of TUBITAK and compares them with the earlier one (2012). The effects of year-to-year changes on the consistency of the ranks of journals are also studied. Note that, as the timeframe is short (2012-2014), we do not intend to study the impact of such changes on the authors' behaviour in terms of which journals they prefer to submit their papers to, journals' acceptance rates or the length of time it takes to publish therein. Rather, we try to understand the motives behind changes along with their effects on journal scores, which in turn determine the rank of each journal and thus the amount of monetary support that TUBITAK provides to the authors of papers that appeared in a specific journal.

TUBITAK's Support Program of International Scholarly Publications

Since 1993, TUBITAK provides monetary support to the authors of scholarly papers that appear in journals indexed by Thomson Reuters as an incentive to increase the number of such publications. The journal impact factor (JIF) was the sole criterion for support until 2013. As is well known, the impact factor (IF) of a journal is measured by the number of citations it gets in a given year to the papers published in it in the previous two years. Thomson Reuters publishes JCRs annually in which journals in each subject discipline covered by Science Citation Index (SCI) and Social Sciences Citation Index (SSCI) are ranked according to their JIFs. TUBITAK used JCRs to determine the eligible journals and

categorized the top 25% of journals in each subject discipline as Group "A", the next 25% of journals as Group "B" and the remaining 50% of journals as Group "C" (and "Group D" for social science journals—the bottom 10% of the remaining 50% of journals) (UBYT Programı, 2012).[1]

In 2013, TUBITAK has almost quadrupled the amount of support per paper. In parallel with this decision, TUBITAK also changed the rules to further classify journals with high IFs by developing its own "journal impact factor". Rather than simply classifying journals as A, B, C, and D on the basis of JCR's two-year JIF data, TUBITAK decided to use JCR's five-year JIFs and cited half-lives of journals in each discipline and multiplied the two figures to come up with its own JIF and ranked journals accordingly. (Cited half-life of a journal is the median—in years—of citations to papers published in it in a given year and depends on how fast the literature obsolesces in subject disciplines.) TUBITAK then took the average TUBITAK JIF of ranked journals and identified the journals with 2 standard deviations (SD) above and below the average to award them the maximum (5,000.00 Turkish Lira[2]) and minimum (500.00 TL) amount of support, respectively. Journals in between were awarded on the basis of a linear transformation formula taking the number of journals in each JCR discipline into account. This formula was criticized by some (Batmaz, 2013) as it happened to downgrade the ranks of some "A class" Archaeology journals considerably, thereby making them least supported ones. Similarly, the 2013 algorithm ranked 56% of Geology journals lower, including *Tectonics*, one of the most prestigious journals in this discipline (Yaltırak, 2014, p. 18).

Apparently, the new algorithm did not fulfill its objectives and TUBITAK, after using it for only one year, quickly replaced it in 2014 with the one that is based on JCR's article influence score. The 2013 transformation formula was used in 2014 to determine the exact amount to be paid to each journal (TUBITAK, 2013; 2014 Yılı, 2014). Comparable to IF, average influence score (AIS) is "a measure of the average influence, per article, of the papers in a journal" (Bergstrom, West & Wiseman, 2008) and is similar to Google's PageRank algorithm in that citations coming from papers in highly cited journals are weighted more heavily (Franceschet, 2010; Arendt, 2010). It is based on the number of citations, nonetheless. AIS is "the most stable indicator across different disciplines" (Franceschet, 2010) and can therefore be used for interdisciplinary comparisons (Arendt, 2010).

The drawbacks of metrics used by TUBITAK (JIF, TUBITAK's own JIF consisting of JCR's five-year IF and cited-half life and AIS) were discussed in detail elsewhere (Tonta, 2014). What follows is a survey based on a sample of 228 journals supported by TUBITAK to see the impact of changes introduced in 2013 and 2014.

Method

In order to find out the impact of most recent changes introduced in 2013 and 2014, we used TUBITAK's list of journals supported in 2012[3] to draw a sample. The list has a total of 11,562 journals. As explained earlier, TUBITAK categorized these journals in 2012 under Groups A, B, C and D according to JIFs reported in Thomson Reuters' JCR. The distribution of 11,562 journals under categories is as follows: Group A: 4,205 (or 36%) journals; Group B: 2,446 (or 21%) journals; Group C: 4,711 (or 41%) journals; and Group D: 200 (or 2%) journals. Social sciences journals constituted about one third of all journals. We selected a sample 232 journals (or 2% of the population) using stratified sampling method. Journals under Groups A, B, C and D formed the four strata. Two numbers between 1 and 100 were

[1] For more detail on TUBITAK's classification of journals, see Tonta (2014).

[2] *Circa* 2,000.00 USD.

[3] Available at http://ulakbim.tubitak.gov.tr/tr/hizmetlerimiz/ubyt-yayin-tesvik-programi.

identified (37 and 54) randomly and every 37th and 54th journal titles were selected. Table 1 provides population parameters and sample statistics.

The distribution of Science and Social science journals in the sample is quite similar to that of population. This can be interpreted as an indication of the generalizability of findings to the population with a calculated margin of error. The original sample size was 232 but 4 journals under Group D were later discarded to simplify the comparisons. Journals supported in 2013 and 2014 are not available as single lists but can be searched using a search engine available at the site.[4] All 228 journal titles in the sample were searched and their journal scores as well as the amount of support they would get were recorded. Six journals[5] in the 2012 list were no longer available in 2013 and 2014 among the supported journals and they were replaced with the next ones (e.g., 38th or 55th record) provided they were in the same category of Science and Social Science journals (e.g., Groups A, B, and C).

Table 1. Population parameters and sample statistics.

	Population parameters						*Sample statistics*					
	Science		*Social Science*		*Total*		*Science*		*Social Science*		*Total*	
Group	*N*	*%*	*N*	*%*	*N*	*%*	*N*	*%*	*N*	*%*	*N*	*%*
A	2037	48	2168	52	4205	100	40	48	44	52	84	100
B	1824	75	622	25	2446	100	36	72	14	28	50	100
C	3763	80	948	20	4711	100	77	82	17	18	94	100
D	--	--	200	100	200	100	--	--	4	100	4	100
Total	7624	100	3938	100	11562		153		79		232	

It should be noted that the minimum and maximum amounts for 2012, 2013 and 2014 were fixed (433.00 TL and 1,300.00 TL for 2012 and 500.00 TL and 5,000.00 TL for 2013 and 2014). As journals in 2012 were awarded fixed amounts of support depending on which group they belonged to, the figure for each journal was obtained by checking its group (e.g., A, B, C) as well as its being a Science or Social science journal. Social science journals were paid twice the amount of what is determined for each group (e.g., the author of a paper published in a Social science journal under group A was awarded 2,600.00 TL instead of 1,300.00 TL).

Findings

Table 2 below provides descriptive statistics for 228 journal titles including the quartiles. Despite the fact that the amount of support was increased in 2013 to 5,000.00 TL, the mean and median values do not seem to be affected much from this increase. The percentage of increase for the journals in the 3rd quartile is noticeable (19%), the reasons for which will be discussed shortly.

Figure 1 provides the scatter graph of the amount of support given by TUBITAK in 2012, 2013 and 2014 to the authors of papers that appeared in 228 journals sampled. Note that the blue line represents the 2012 figures and ranked in descending order by the amount of support. The amount was fixed depending on which group the journal belonged to. The authors of articles that appeared in Groups A, B, and C journals were paid 1,300.00, 867.00, and 433.00 Turkish Lira (TL), respectively.[6] If the paper appeared in a Social science journal,

[4] http:// http://www.ulakbim.gov.tr/

[5] Or, they might have been discontinued or their names might have changed. Replaced journal titles are: *Journal of Dental Research, Tulsa Studies in Women's Literature, Journal of Electronic Imaging, Plasma Physics Reports, and Vie et Milieu – Life and Environment*.

[6] The authors of case studies, technical communications, letters to the editors, etc. received half this amount.

the amount of support is doubled so that the authors of Social science papers will be further encouraged. Therefore, the solid blue line at 2,600.00 TL and 1,733.00 TL represent both 43 Group A and 14 Group B Social science journals, respectively, whereas the blue line at 1,300.00 TL represents 41 Group A Science journals. The 867.00 TL band represents both 35 Group B Science journals and 17 Group C Social science journals. The 433.00 TL band represents 78 Group C Science journals.

Table 2. The amount of support (in Turkish Lira*).

	2012	*2013*	*2014*	*Increase 2013-2014 (%)*
Mean	1176	1317	1403	7
Minimum	433	500	500	0
1st quartile	433	533	558	5
Median	867	829	874	5
3rd quartile	1408	1518	1806	19
Maximum	2600	5000	5000	0

*Rounded to the nearest whole number.

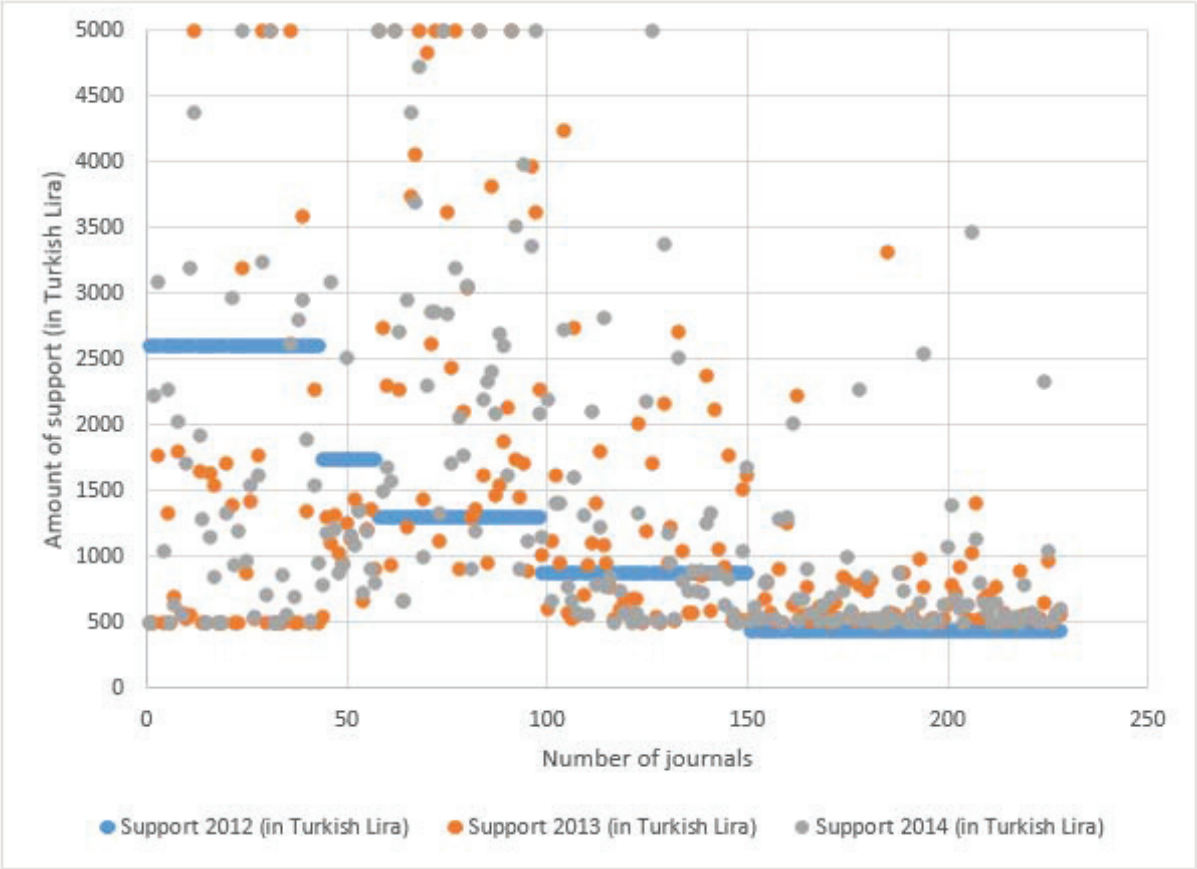


Figure 1. The scatter of journals by the amount of support in 2012, 2013 and 2014 (N = 228).

As indicated earlier, the maximum amount of support in 2013 was increased to 5,000.00 TL (the minimum being 500.00 TL). Note that the Group A journals of 2012 received relatively less support in 2013 and 2014. Out of 84 journals classified under Group A in 2012, only 15 (18%) maintained their top positions in the following years.[7] However, the positions of Social

[7] The amount between 500.00 TL and 5,000.00 TL was divided into three equal groups and the ones that were awarded between 3,500.00 TL and 5,000.00 TL are considered as top journals.

science journals classified under Group A fluctuated more than that of Science journals. Only 3 out of 43 Social science journals (7%) maintained their top positions as opposed to 12 out of 41 Science journals (29%).

Note that 2013 and 2014 figures are scattered without seemingly any discernible pattern (Fig. 1), as the 2012 figures are ranked in descending order by the amount of support and they do not necessarily correspond with the amounts in 2013 and 2014. Although statistically significant, the correlation between the amount of support to journals in 2012 and 2013 and that in 2012 and 2014 was rather low (Pearson's r = .289 and .231, p = .000, respectively). The correlation between the 2013 and 2014 journals was moderate (Pearson's r = .767, p = .000) (see Fig. 2).

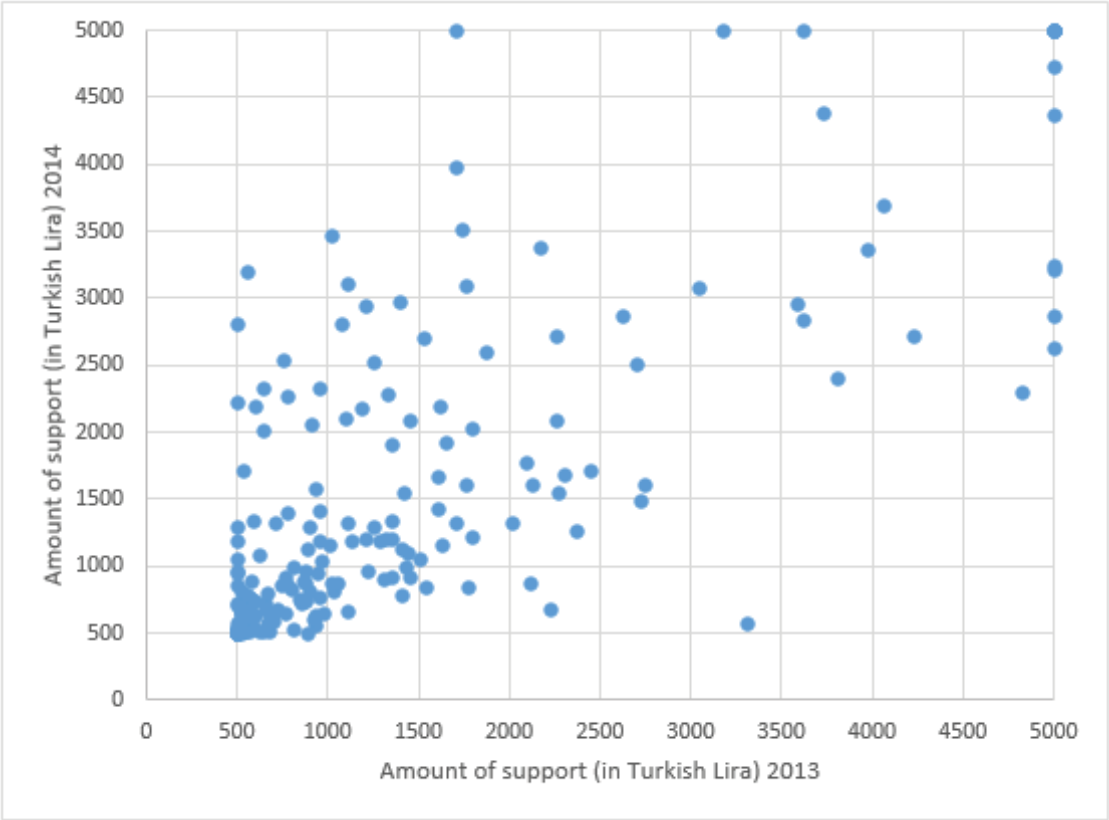


Figure 2. The scatter of journals by the amount of support in 2013 and 2014 (N = 228).

It is estimated that some 30,000 scholarly journals are published in the world. Thomson Reuters indexes about 12,000 of them and TUBITAK supports almost all of them (TUBITAK's 2012 journal list had 11,562 journal titles). It should be pointed out that TUBITAK's threshold for support is rather low. As Figures 3 and 4 below show, about one third of journals barely meet the minimum criteria and get the minimum amount of support (500.00 TL). It is reasonable to suggest that after careful consideration support to more than 3,000 journals can easily be discontinued.

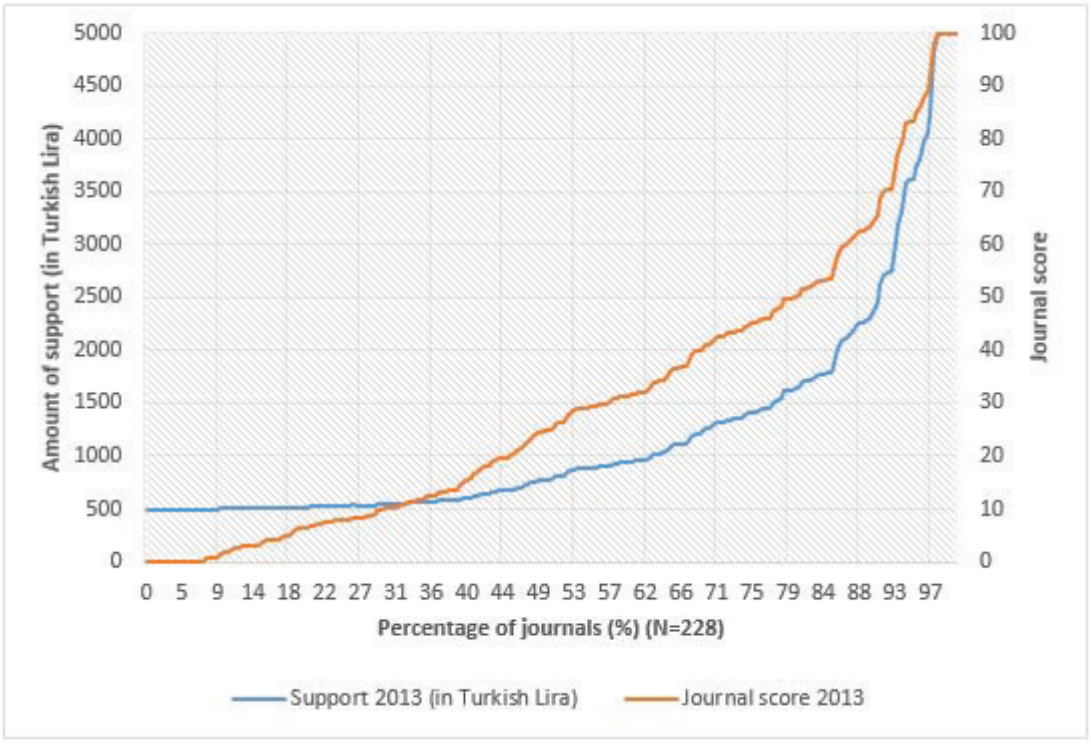


Figure 3. Relationship between journal score and the amount of support in 2013.

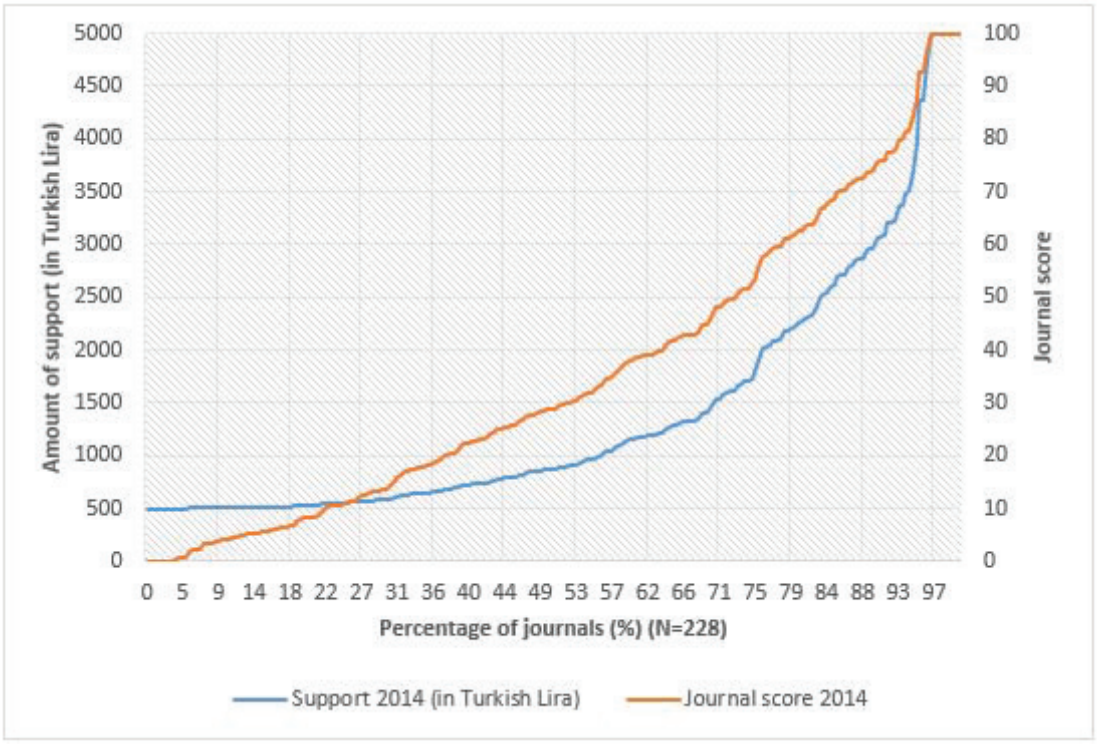


Figure 4. Relationship between journal score and the amount of support in 2014.

It should also be pointed out that the new policy discourages the authors of papers that appear in journals with low Article Influence Scores to seek support. As Figure 3 and 4 show, the gap between the journal scores and the amount of support starting from about 27%-35% gets widened. In other words, the amount of support is not that high for journals with relatively

lower AISs. More than 90% and 80% of journals received less than 2,500.00 TL (half the full amount of 5,000.00 TL) in 2013 and in 2014, respectively. Journals that received more than 4,000.00 TL support were about 5% of all journals in both 2013 and 2014. The situation was even worse for Social science journals (Fig. 5). This trend can also be followed from the last column of Table 2. The percentage of increase for the journals in the third quartile between 2013 and 2014 was 19% while it was only 5% for the journals in the first and second quartiles. This could be interpreted as a positive sign to encourage the authors to publish in more prestigious journals with higher AISs. Note that if the amount was less than 100.00 TL per co-author for papers with multiple authors, no support is provided. This is a further disincentive for authors not to claim the TUBITAK support for papers that appear in journals with low impact factors or article influence scores.

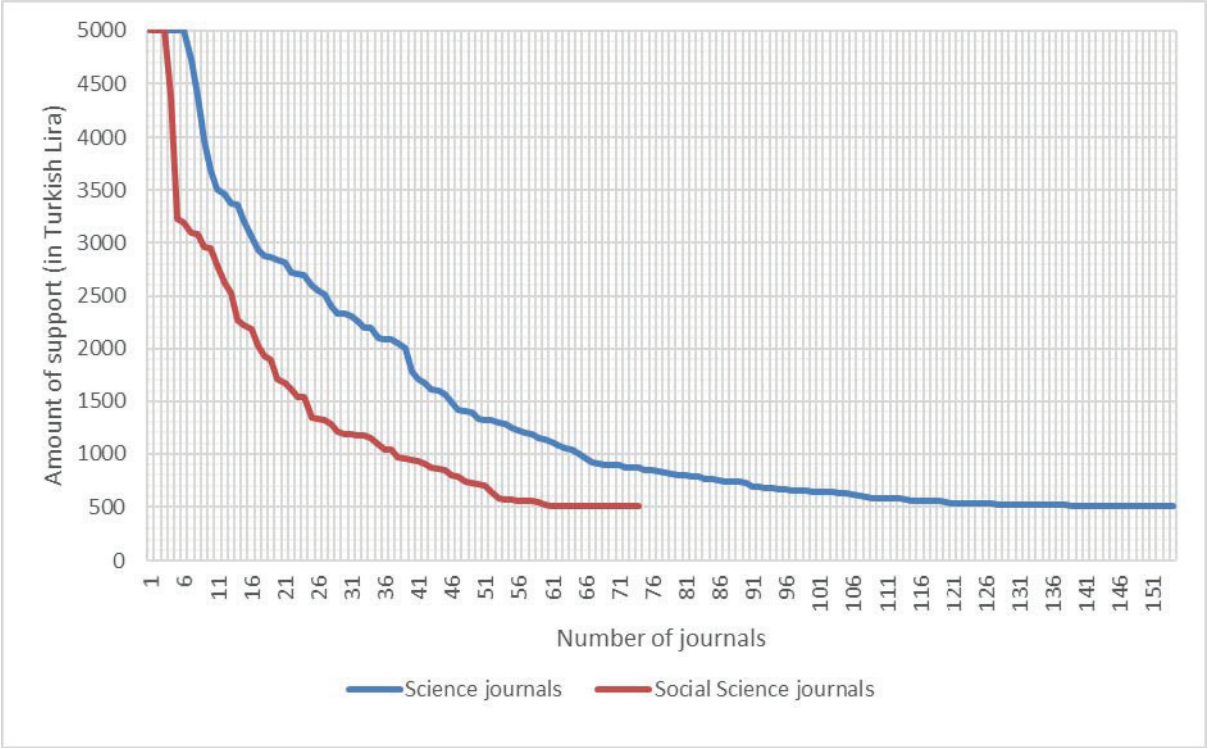


Figure 5. The amount of TUBITAK support for Science and Social science journals in 2014.

As we explained earlier, TUBITAK classified the second half of journals in Science disciplines listed in JCR under Group "C" and provided minimum support (433.00 TL per article) for these journals. (For Social Science disciplines, the second half of journals were divided into two: the top 40% of them being labeled as Group "C" and the remaining 10% as Group "D". Later, TUBITAK stopped supporting the authors of papers publishing in journals under Group "C" in Sciences (i.e., the last 50% of journals) and Group "D" in Social Sciences (i.e., the last 10% of journals) (UBYT Uygulama, 2012). As Group C Science journals constituted about one third of all journals supported in 2012, we wanted to see if they get supported after the policy changes in 2013 and 2014. Our sample included 77 Group C Science journals (one third of all sampled journals) (Table 1). It appears that all of them got supported both in 2013 and 2014. However, the overwhelming majority of them received very little support. As mentioned earlier, the 2013 algorithm was based on five-year JIFs and cited half-lives whereas the 2014 algorithm was based on article influence scores. Recall that the amount of support was increased almost four times starting from 2013. If TUBITAK were to continue supporting Group C Science journals, the amount would have been equal to 1,665.00 TL. Yet, the number of Group C Science journals receiving 1,665.00 TL (or higher) support

was only 2 in 2013 and 5 in 2014. The average amount of support in 2013 and 2014 were 701.00 TL (median=564.00 TL) and 770.00 TL (median=577.00 TL), respectively.

As JIFs and article influence scores are both based on the number of citations, it is not that surprising to see that journals that performed poorly in 2012 did so, too, in 2013 and 2014. What is surprising to see though is that TUBITAK seems to have nullified its earlier decision of not supporting Group C Science journals. A very few of those journals performed differently in 2013 and 2014 when new algorithms were used.

Discussion and Conclusion

It appears that the two algorithms used by TUBITAK in 2013 and 2014 are not that different from each other after all, even though the former was based on Thomson Reuters' JIFs and cited half-lives and the latter on article influence scores (AIS). However, as mentioned earlier, AIS is the most stable indicator and the average influence of journals can therefore be comparable across disciplines (Franceschet, 2010; Arendt, 2010). JIFs and AISs are highly correlated with each other and papers published in high impact journals usually have high AISs (Arendt, 2010; Rousseau & STIMULATE 8 Group, 2009). Arendt (2010) examined the relationship between the two metrics using 5,900 journals listed in JCR Science Edition (2007) and found that both JIFs and AISs vary by discipline. Moreover, the correlation between the two metrics was quite high (Pearson's r (172) = .896) and statistically significant ($p < .001$). Arendt (2010) cautioned that these two metrics should not be used formulaically for research assessment and for ranking scientific papers, authors or institutions.

This advice should be taken into account by TUBITAK as well. As the algorithm based on AIS is more stable and does not vary that much by scientific disciplines (Arendt, 2010; Franceschet, 2010), its use should be monitored closely by TUBITAK to see if it merits further refinement.

The support to journals in the lower end of the scale should be discontinued. Having decided in 2012 to discontinue support to Group C Science journals, it is not clear why TUBITAK reversed its decision the following year without monitoring how these journals performed with the new algorithms used in 2013 and 2014. In fact, the performance of all journals should be monitored to fine-tune the algorithms used.

TUBITAK is of the opinion that its support program caused to increase the number of scientific publications over the years. Turkey has indeed performed very well and became the 18th country in the world in terms of the number of scholarly papers published in ISI-indexed journals. However, the positive correlation between the amount of support provided by TUBITAK and the number of papers with Turkish affiliations is not a strong argument in and of itself[8] to justify the continuance of the support program because correlation does not necessarily mean causation. The existing support to papers published in low impact journals could very well be the main cause of this positive correlation. This merits further research because TUBITAK support does not seem to have encouraged the authors to publish in more prestigious journals.

In conclusion, bibliometric performance measures alone are not the sole criteria for research assessment and, as the Board of Directors of IEEE recently recommended, they "**should be applied only as a collective group (and not individually)**" (IEEE, 2013, original emphasis).

References

2014 yılı UBYT Programı teşvik miktarları hesaplama yöntemine dair bilgi notu (A note on the calculation of the amounts of support in TUBITAK's UBYT Program of 2014). (2014). Retrieved, January 25, 2015, from http://ulakbim.tubitak.gov.tr/sites/images/Ulakbim/ubyt_2014_hesap.pdf.

[8] The number of universities and researchers in Turkey have also increased tremendously during this period.

Arendt, J. (2010). Are article influence scores comparable across scientific fields? *Issues in Science and Technology Librarianship*, No. 60. Retrieved, January 25, 2015, from http://www.istl.org/10-winter/refereed2.html.

Balaram, P. (2013, May 25). Research assessment: Declaring war on the impact factor. *Current Science*, *14*(10), 1267-1268.

Batmaz, A. (2013, June 14). Türkiye'de bilim üretimi ve arkeoloji (Science production in Turkey and Archaeology). *Cumhuriyet Bilim ve Teknoloji*, (1369), 18. Retrieved, January 25, 2015, from http://www.arkeolojikhaber.com/?p=2569.

Bergstrom, C.T., West, J.D., & Wiseman, M.A: (2008). The Eigenfactor™ metrics. *The Journal of Neuroscience*, 28(45): 11433-11434. Retrieved, January 26, 2015, from http://www.jevinwest.org/Documents/Bergstrom_J_neurosci_2008.pdf

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. Retrieved, January 26, 2015, from doi:10.1371/journal.pone.0006022.

Browman, H.I. & Stergiou, K.I. (2008). Factors and indices are one thing, deciding who is scholarly, why they are scholarly, and the relative value of their scholarship is something else entirely (editorial). *Ethics in Science and Environmental Politics*, *8*, 1-3. Retrieved, January 26, 2015, from http://www.int-res.com/articles/esep2008/8/e008p001.pdf.

Casadevall, A. & Fang, F.C. (2014). Causes for the persistence of impact factor mania. *mBio*, 5(2). Retrieved, June 25, 2014, from http://mbio.asm.org/content/5/2/e00064-14.full.pdf.

Franceschet, M. (2010). Journal influence factors. *Journal of Informetrics*, 4(3), 239-248. Retrieved, January 26, 2015, from https://users.dimi.uniud.it/~massimo.franceschet/publications/joi10b.pdf

Guerrero, R. (2001, August). Misuse and abuse of journal impact factors. *European Science Editing*, 27(3): 58-59.

Hecht, F., Hecht, B.K. & Sandberg, A.A. (1998, July 15). The journal "impact factor": A misnamed, misleading, misused measure. *Cancer Genetics*, *104*(2), 77-81.

IEEE. (2013, September 9). Appropriate use of bibliometric indicators for the assessment of journals, research proposals, and individuals. Retrieved, April 7, 2015, from http://www.ieee.org/publications_standards/publications/rights/ieee_bibliometric_statement_sept_2013.pdf

Jawaid, S.A. (2014). Despite misuse and abuse, journal impact factor will retain its impact and won't fade away soon (editorial). *Journal of Postgraduate Medical Institute*, *28*(1), 1-4.

Kotur, P.F. (2013, August 10). Impact factor – the misnamed, misleading and misused measure of scientific literature. *Current Science*, *105*(3), 289-290. Retrieved, January 26, 2015, from http://www.currentscience.ac.in/Volumes/105/03/0289.pdf.

Lawrence, P.A. (2008). Lost in publication: how measurement harms science. *Ethics in Science and Environmental Politics*, *8*, 9-11. Retrieved, January 26, 2015, from http://www.int-res.com/articles/esep2008/8/e008p009.pdf.

Marks, M.S., Marsh, M., Schroer, T.A., & Stevens, T.H. (2013, June). Misuse of journal impact factors in scientific assessment (editorial). *Traffic*, *14*(6), 611-612. Retrieved, January 26, 2015, from http://onlinelibrary.wiley.com/doi/10.1111/tra.12075/full.

Marx, W. & Bornmann, L. (2013). Journal Impact Factor: "the poor man's citation analysis" and alternative approaches. *European Science Editing*, 39(2), 62-63. Retrieved, January 25, 2015, from http://www.ease.org.uk/sites/default/files/aug13pageslowres.pdf.

San Francisco Declaration on Research Assessment: Putting science into the assessment of research. (2012, December 16). Retrieved, January 25, 2015, from http://am.ascb.org/dora/files/SFDeclarationFINAL.pdf.

Simons, K. (2008, October 10). The misused impact factor. *Science*, 322, 165. Retrieved, January 26, 2015, from https://java-srv1.mpi-cbg.de/publications/getDocument.html?id=8a8182da238c39d10123955066a00100.

Rousseau, R. & STIMULATE 8 Group. (2009). On the relation between the WoS impact factor, the eigenfactor, the SCImago journal rank, the article influence score and the journal index. (Technical report). Retrieved, January 26, 2015, from http://eprints.rclis.org/16448.

Seglen, P.O. (1997). Why impact factor of journals should not be used for evaluating research. *British Medical Journal*, 314, 497-502.

TÜBİTAK Türkiye Adresli Uluslararası Bilimsel Yayınları Teşvik Programı Uygulama Esasları (TUBITAK's Principles to Support International Scientific Publications with Turkish Affiliations). (2013). Retrieved, January 25, 2015, from http://www.tubitak.gov.tr/sites/default/files/esaslar_v_2_vers.2_2.pdf.

Todd, P.A., & Ladle, R.J. (2008). Hidden dangers of a 'citation culture'. *Ethics in Science and Environmental Politics*, *8*(1), 13-16. Retrieved, January 26, 2015, from http://www.int-res.com/articles/esep2008/8/e008p013.pdf.

Tonta, Y. (2014). Use and misuse of bibliometric measures for assessment of academic performance, tenure and publication support. *Metrics 2014: Workshop on Informetric and Scientometric Research (SIG/MET). 77th Annual Meeting of the Association for Information Science and Technology, October 31-November 5, 2014, Seattle, WA*. Retrieved, January 26, 2015, from http://yunus.hacettepe.edu.tr/~tonta/yayinlar/tonta-asist2014-seattle-sig-met-misuse-of-bibliometric-indicators.pdf

Yaltırak, C. (2014, June 21). TÜBİTAK yayın teşvik sistemini değiştirmeli! (TUBITAK has to change its publication support system!) *Cumhuriyet Bilim ve Teknoloji*, (1409), 18.

What's Special about Book Editors? A Bibliometric Comparison of Book Editors and other Flemish Researchers in the Social Sciences and Humanities

Truyken L.B. Ossenblok[1] and Mike Thelwall[2]

[1] *Truyken.Ossenblok@uantwerpen.be*
Centre for Research & Development Monitoring (ECOOM), Faculty of Political and Social sciences, University of Antwerp, Middelheimlaan 1, 2020 Antwerp (Belgium) (corresponding author)

[2] *M.Thelwall@wlv.ac.uk*
Statistical Cybermetrics Research Group, Faculty of Science and Engineering, University of Wolverhampton (United Kingdom)

Abstract
This paper examines the bibliometric characteristics of book editors and non-editors, focussing on gender, career stage, number of publications and collaboration practices. The data consist of 8970 Flemish affiliated researchers with at least one publication between 2000 and 2011 in the comprehensive Flemish academic bibliometric database (VABB-SHW). The analysis shows that most book editors are established male researchers while most non-editors are non-established male researchers. Moreover, males are more likely to be editors than are females. Half of the established editors edit more than 1 book, in contrast to only a small number of non-established editors. Overall, book editors publish more than non-editors, but, when controlling for career stage, book editors publish even more book chapters and monographs than do non-editors. Although editors are highly collaborative while editing a book, no significant differences were found in the number of collaborative articles, monographs, book chapters and proceedings written by editors and non-editors.

Conference Topic
Country-level studies

Introduction
Bibliometric studies have demonstrated the importance of books to many disciplines belonging to the Social Sciences and Humanities (SSH). There is a growing consensus among researchers and policy-makers that scholarly publication patterns and their underlying research cultures cannot be adequately analyzed without the inclusion of books (Hicks, 2004; Nederhof, 2006; Sivertsen, 2009). So far, this insight has resulted in a limited number of studies on books in the SSH, mostly focused on scholarly monographs. A book publication type that has received far less attention is the edited book. Editing a book often appears to be undervalued for academic careers (Edwards, 2012) but, in Flanders, from 2010 onwards, edited books are included in the funding system (Ossenblok & Engels, 2015) which gives incentives to individual researchers to take on book editorships (Gläser & Laudel, 2007).

We define an edited book here as a collection of chapters written by different authors, gathered and harmonized by one or more editors (Ossenblok & Engels, 2015) and identifiable by the presence of an ISBN. Edited books have been shown to comprise a sizeable share of the publication output of many SSH disciplines, especially in the humanities (Leydesdorff & Felt, 2012; Nederhof, 2006). In Flanders, the Northern Dutch-speaking part of Belgium, about 2% of all peer reviewed publications in the SSH are edited books, with up to 6% in Linguistics, Literature and Theology (Engels, Ossenblok, & Spruyt, 2012). Compared to monographs, edited books have significantly higher citation rates, especially in social science disciplines (Torres-Salinas, Robinson-Garcia, Cabezas-Clavijo, & Jiménez-Contreras, 2013).

This paper presents a bibliometric case study of the characteristics of book editors, for which, to the best of our knowledge, no previous studies exist. We analyse comprehensive

publication data and present four elements of a general profile of these scholars: career stage; gender; number of publications; and collaboration practices. We hypothesise that scholars tend to edit books only when they are established researchers that are at the forefront of scholarly collaboration.

Data and methods

The data set consists of 8970 authors affiliated with one of the five Flemish universities and who have published a minimum of one peer reviewed publication in the period 2000-2011: a journal article, monograph, edited book, book chapter and/or proceedings paper included in the VABB-SHW (for a full account see: Engels et al., 2012). Because of the use of this database for funding in Flanders, this database appears to be close to exhaustive in its coverage of Flemish research. In addition to the data found in the VABB-SHW, we also determined the gender of all authors. For this, two researchers independently divided all unambiguous first names into two groups: male names and female names. The remaining authors were looked up on the internet, resulting in an additional 1462 gender matches.

A comparison was made between two subsets: book editors (researchers who have published a minimum of 1 peer reviewed edited book in the period under study); and all other researchers, called here non-editors although they may be journal editors or may have edited books during other periods of time. Furthermore, we differentiated between established and non-established researchers. Established researchers are defined in this study as having a total of 12 publications or more and at least one publication in a minimum of 6 different years in the period 2000-2011. These heuristics were chosen after inspection of typical properties of authors in the database. Of course, non-established researchers may have many publications within up to five years, may have a prolific consistent set of outputs before or after the period analysed, or may have many outputs of a type not recorded in the database (e.g., book reviews, performances). Nevertheless, the criteria seem to be effective at differentiating between two sets of researchers, the first of which contains researchers that can reasonably be thought of as being established and the second of which probably contains a much lower proportion of established researchers. Cramer's V was used to measure the strength of the correlation between the different subsets, resulting in a number between 0 (no association) and 1 (maximum association). In addition the Mann-Whitney U test, a rank-based nonparametric test, was used to determine whether there were differences between the subsets on the different characteristics under study, using $p=0.05$ as the threshold for statistical significance.

Results

Career stage and gender

Figure 1 shows the proportion and number of established and non-established, male and female editors and non-editors in our study. In total, 676 (7.5%) researchers had published one or more edited books (i.e., editors), and 8970 (92.5%) researchers had not published an edited book (i.e., non-editors). Figure 1 demonstrates that 55.9% (n=378) of editors are established researchers whereas 13.3% (n=1102) of non-editors are established researchers. Furthermore, 74.3% (n=502) of editors are male whereas to 58.9% (n=4883) of non-editors are male. In addition, 9.3% of all male researchers are editors and 4.9% of all female researchers are editors. Furthermore, 25.5% of all established researchers are editors, whereas only 4% of all non-established researchers are editors. Altogether, 43.5% (n=294) are male established editors, 30.8% (n=208) are male non-established editors, 13.3% (n=90) are female non-established editors and 12.4% (n=84) are female established editors. Different proportions occur in the subgroup of the non-editors where 49.1% (n=4070) are male non-

established researchers, 37.6% (n=3122) are female non-established researchers, 9.8% (n=813) are male established researchers and 3.5% (n=289) are female established researchers.

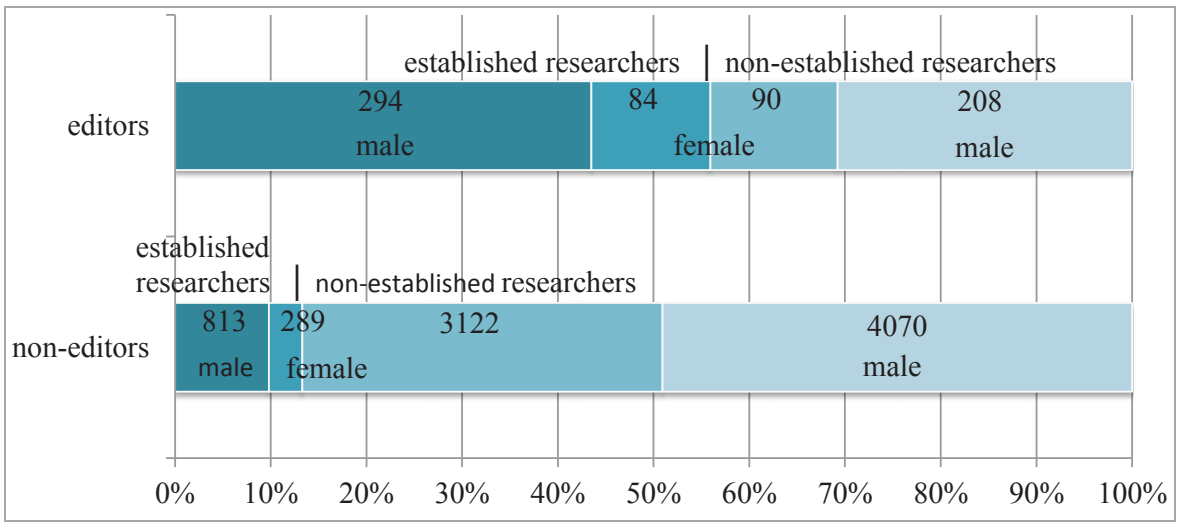


Figure 1: Share and number of established and non-established, male and female editors and non-editors (2000-2011).

There is a moderate association (Cramer's V=0.134; p=.000) between gender and career status overall (see also Figure 1). However, when looking at the different subsets, the correlation between gender and career status is stronger within the subset of non-editors (Cramer's V=0.119; p=.000) than within the subset of editors (Cramer's V=0.091; p=.000). Overall, though, career status has a stronger association with editorship than with gender (resp. Cramer's V=0.304; p=.000 and Cramer's V=0.083; p=.000). Therefore in the rest of this study we will focus on differences in career status rather than gender.

Number of publications

Table 1 shows the mean and median number of edited books, articles, book chapters, monographs and proceedings for all editors and non-editors. In addition, the table displays the difference between non-established and established researchers. Overall, editors publish on average a greater number of all publication types than do non-editors. However, established non-editors publish on average more articles than do established editors. Mann-Whitney U tests were run to test for differences in numbers of publications between editors and non-editors for all publication types except edited books. The distributions of all the publication types for editors and non-editors and for established and non-established researchers were visually similar. The differences between editors and non-editors are statistically significant for all publication types (all p=.000). When comparing established editors and established non-editors, all differences are significantly different (p=.000) except for the numbers of proceedings (p=.138). When comparing non-established editors with non-established non-editors, the differences for articles (p=.119) and proceedings (p=.911) were not significantly different, whereas the differences for book chapters and monographs were (both p=.000).

Furthermore, Table 1 shows that the median of numbers of edited books differ between established and non-established editors. Non-established editors are more likely to have (co−)edited one book whereas established editors are more likely to have more than 1 edited book. More specifically, 83.2% of all non-established editors have one edited book, whereas 48.4% of all established editors have one edited book, 24.3% have two edited books and 27.2% have three or more edited books.

Table 1: The mean and median (med) number of edited books, articles, book chapters, monographs and proceedings for all established and non-established editors and non-editors (2000-2011).

		edited books		*articles*		*book chapters*		*monographs*		*proceedings*	
		mean	*med*	*mean*	*med*	*mean*	*med*	*mean*	*med*	*mean*	*med*
Editor	established researcher	2.17	2	20.62	14	7.92	6	0.59	0	0.97	0
	non-established researcher	1.22	1	2.93	2	2.31	2	0.16	0	0.17	0
	total	1.76	1	12.82	7	5.44	4	0.40	0	0.62	0
non-editor	established researcher	-	-	26.00	18	1.57	1	0.22	0	0.82	0
	non-established researcher	-	-	3.00	2	0.29	0	0.03	0	0.16	0
	total	-	-	6.06	2	0.46	0	0.05	0	0.24	0

Collaboration practices

For both editors and non-editors, Figure 2 shows the proportion of their edited books, articles, book chapters, monographs and proceedings that have been published in collaboration (i.e., multiple authored versus single authored publications). Editors collaborate the most while editing a book (90.3%; n=1827), which is in agreement with previous research demonstrating that most edited books are co-edited (Ossenblok & Engels, 2015). Furthermore, established editors collaborate more than non-established editors for all publication types under study (p=.000). Altogether, though, non-editors seem to collaborate more for articles, book chapters, monographs and proceedings than do editors. Mann-Whitney U tests were run to determine if editors and non-editors differ significantly in their numbers of collaborative publications. The different distributions of all the publication types, except edited books, were visually similar. The numbers of collaborative publications of editors and non-editors were statistically significantly different for book chapters and monographs (both p=.000) but not for articles (p=.282) and proceedings (p=.116). Thus, non-editors collaborate significantly more in book chapters and in monographs than do editors. In addition, when comparing non-established editors with non-established non-editors, no significant difference in the number of collaborative publications was found for all publication types separately (but p=.000 for articles, monographs and book chapters; p=.005 for proceedings). However, when distinguishing between established editors and non-editors, the differences are significant for all publication types separately (p=.000) except for proceedings (p=.208). In sum, established non-editors collaborate more than do established editors for articles, monographs and book chapters.

Discussion and conclusions

Within a comprehensive collection of Flemish affiliated authors' publications for 2000-2011, this paper demonstrates that 7.5% of the authors have edited one or more books, that more than half of the book editors are established researchers, and that 3 in 4 editors are male.

Female researchers are less likely to be established than are male researchers and this difference is more pronounced for non-editor than for editors. As career status in this study is defined through numbers of publications and publication years, these findings confirm previous findings that male researchers are often more productive than are their female colleagues (Larivière et al., 2013; Puuska, 2010).

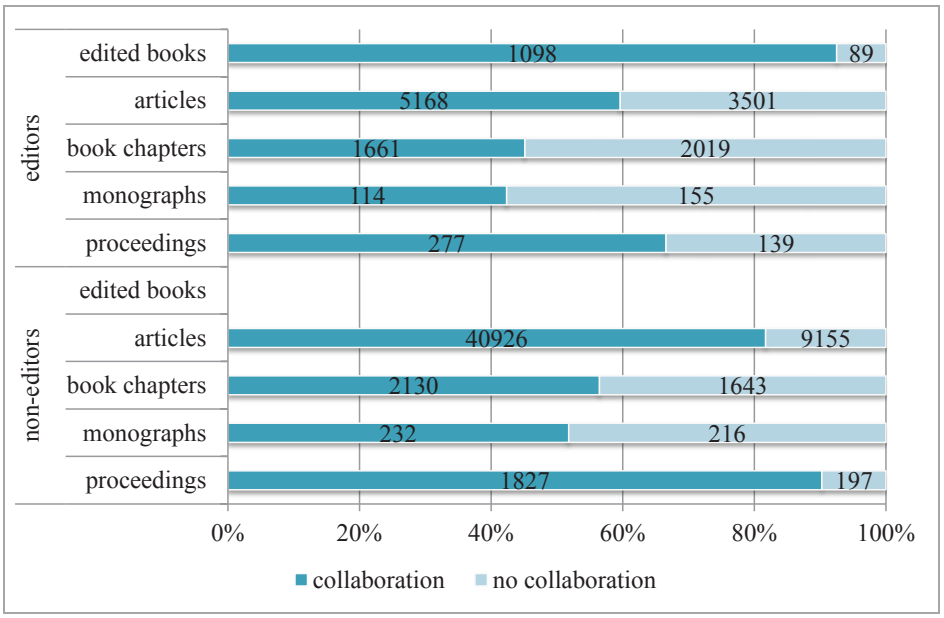


Figure 2: The proportion of collaborative and solo publications for all editors and non-editors by publication type.

Editors tend to publish significantly more articles, book chapters, monographs and proceedings than do non-editors. However, the differences are not statistically significant between the average number of proceedings of established editors and non-editors and between the average number of articles and proceedings of non-established editors and non-editors. Most non-established editors published only 1 edited book in the period under study, whereas more than half of the established editors published 2 or more edited books. This might be due to the need for a large network and good networking skills for gathering contributions from individual chapter authors for an edited book (Edwards, 2012; Thomas & Hrebenar, 1993). We therefore expected editors to be more collaborative than were non-editors for all publication types, but although 9 out of 10 editors collaborated while editing a book, non-editors collaborated significantly more for book chapters and monographs than did editors. Furthermore, no significant difference was found in the number of collaborative articles and proceedings between editors and non-editors. As edited books are more common in humanities disciplines (Engels et al., 2012) and the humanities have been known to collaborate less than the social sciences in articles and book chapters (Ossenblok, Verleysen, & Engels, 2014), the low level of collaboration of editors might be due to them tending to be humanities scholars.

Overall, the findings offer a first insight into some of the bibliometric characteristics of editorship. Future research will focus on disciplinary differences in collaboration practices between book editors and non-editors. A more detailed analysis of collaboration practices will involve not only the number of collaborative publications, but also the number of co-authors. As previous research (Ossenblok & Engels, 2015) has shown, edited books are often published in English, and so the study of the number of international co-authors and co-editors will broaden our knowledge about the international nature of the collaboration network of the editors. In addition, links between book editors and their chapter authors

would provide a more complete picture of the collaboration practices of book editors. This would contribute greatly to our understanding of collaborative practices in the SSH.

Acknowledgments

The authors thank their colleagues Nele Dexters, Tim Engels, Raf Guns and Frederik Verleysen for their useful comments.

References

Edwards, L. (2012). Editing academic books in the humanities and social sciences: Maximizing impact for effort. *Journal of Scholarly Publishing, 44,* 61-74.

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.

Gläser, J. & Laudel, G. (2007). Evaluation without evaluators. In R.Whitley & J. Gläser (Eds.), *The changing governance of the sciences. The advent of research evaluation systems* (pp. 127-151). Dordrecht: Springer Science.

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.

Larivière, V., Ni, C., Gingras, Y., Cronin, B., & Sugimoto, C. R. (2013). Global gender disparities in science. *Nature, 504,* 211-213.

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,* 28-34.

Nederhof, A. J. (2006). Bibliometric monitoring of research performance in the social sciences and the humanities: A review. *Scientometrics, 66,* 81-100.

Ossenblok, T. L. B. & Engels, T. C. E. (2015). Edited books in the social sciences and humanities: Characteristics and collaboration analysis. *Scientometrics, Under review*.

Ossenblok, T. L. B., Verleysen, F. T., & Engels, T. C. E. (2014). Co-authorship of journal articles and book chapters in the social sciences and humanities (2000-2010). *Journal of the American Society for Information Science & Technology, 65,* 882-897.

Puuska, H.-M. (2010). Effects of scholar's gender and professional position on publishing productivity in different publication types. Analysis of a Finnish university. *Scientometrics, 82,* 419-437.

Sivertsen, G. (2009). Publication patterns in all fields. In F.Aström, R. Danell, B. Larsen, & J. W. Schneider (Eds.), *Celebrating scholarly communication studies: A Festschrift for Olle Persson at his 60th birthday* (pp. 55-60). ISSI.

Thomas, C. S. & Hrebenar, R. J. (1993). Editing multiauthor books in political science: Plotting your way through an academic minefield. *Political Science and Politics, 26,* 778-783.

Torres-Salinas, D., Robinson-Garcia, N., Cabezas-Clavijo, Á., & Jiménez-Contreras, E. (2013). Analyzing the citation characteristics of books: Edited books, book series and publisher types in the Book Citation Index. *Scientometrics, 98,* 2113-2127.

Scientific Cooperation in the Republics of Former Yugoslavia Before, During and After the Yugoslav Wars

Dragan Ivanović[1], Miloš M. Jovanović[2] and Frank Fritsche[3]

[1] *dragan.ivanovic@uns.ac.rs*
University of Novi Sad, Faculty of Technical Sciences, Trg Dositeja Obradovića 6, Novi Sad (Serbia)

[2] *milos.jovanovic@int.fraunhofer.de*
Fraunhofer Institute for Technological Trend Analysis, Appelsgarten 2, 53879 Euskirchen (Germany)

[3] *frank.fritsche@int.fraunhofer.de*
Fraunhofer Institute for Technological Trend Analysis, Appelsgarten 2, 53879 Euskirchen (Germany)

Abstract
This paper presents an analysis of scientific research output of the republics of former Yugoslavia for the period 1970-2014. Thomson Reuters' Web of Science (WoS) database was used for data acquisition and 223 135 publications have been analyzed. The Yugoslav Wars were ethnic conflicts fought from 1991 to 1999 on the territory of former Yugoslavia, which accompanied the breakup of the country, and today, each republic of former Yugoslavia is an independent country, as well as the province of Kosovo. Results of the analysis are represented by four figures depicting cooperation networks between former Yugoslav republics and the province of Kosovo for the periods before the Yugoslav wars (from 1970 until 1990), during the wars (from 1991 until 1999), in the first decade after the wars (from 2000 until 2009), and in the last 5 years (from 2010 until 2014). The impact of the wars on scientific cooperation in the republics has been studied.

Conference Topic
Country-level studies

Introduction
The Socialist Federal Republic of Yugoslavia (SFRY) was established in 1946, after World War II. It was divided into six Republics (Serbia, Croatia, Slovenia, Bosnia & Herzegovina, Macedonia and Montenegro) and two autonomous provinces on the north and south of Serbia (Vojvodina and Kosovo). The Yugoslav Wars were ethnic conflicts fought from 1991 to 1999 on the territory of SFRY, which accompanied the breakup of the country. Today, each republic of former SFRY is an independent country. A Kosovo declaration of independence was adopted on 17 February 2008 by the Assembly of Kosovo, but the legality of this declaration have been disputed by the Serbian Government and other countries (e.g. the Russian Federation and China). This paper analyses the scientific cooperation in the republics of former SFRY and the province of Kosovo before, during and after the Yugoslav wars. The purpose of this analysis is to answer how the Yugoslav wars and social crises during and around those wars affected scientific productivity and scientific cooperation in these republics and whether this cooperation has recovered 15 years after the wars.

Related work
Bibliometric analysis is a useful method for characterising scientific research (Moravcsik, 1985; Fu & Ho, 2013) and this method can be used for analysing scientific cooperation in different countries and regions (Leta & Chaimovich, 2002; Wagner & Leydesdorff, 2005; Ho et al., 2010). Citations of a publication are not a direct measure of quality and significance, but they reflect the visibility and impact of the publication on the scientific community (Furlan & Fehlings, 2006; Baltussen & Kindler, 2004). The number of times an article was cited correlates significantly with the number of authors and the number of institutions

involved in collaboration (Figg et al., 2006) and highly cited articles are usually authored by a large number of scientists, often involving international collaboration (Aksnes, 2003). Thus, scientific cooperation is important for the further development of world science and for the further economic development of a region or country.

The impact of social aspects, economic and social crises, political crises and wars on scientific cooperation in some regions has already been studied. For example, de Bruin and colleagues (1991) stated that the cooperation between the Gulf States and former western and eastern bloc has been strongly affected by political crises, which culminated in the Operation Desert Storm in 1990. There are also studies that deal with the countries of the former SFRY like Lewison and Igic (1999), Igic (2002), Lukenda (2006), Đukić et al. (2011) and Kutlača et al. (2015). Furthermore, Jovanović et al. (2010). analysed the publications and cooperation between the republics of former SFRY and the province of Kosovo is analyzed for the years from 1970 until 2007. The authors found that the Yugoslav wars had a severe impact on the cooperation networks of former SFRY republics. Furthermore, they also found that the process of recovery started with the ending of the conflicts, but that scientific cooperation recovered faster in some of those republics. The current paper revisits the data and methods of this study by analysing publications of former SFRY republics and the province of Kosovo from 1970 until 2014, thus broadly extending the database and improving the methodology. Thus, the purpose of this analysis is to answer whether scientific cooperation in all former SFRY republics is fully repaired 15 years after the Yugoslav wars or whether the interpretation of the findings of the 2010 study has to be reformulated.

Methodology

Similar to the 2010 study, Thomson Reuters' Web of Science (WoS) database was used for data acquisition. This time, however, the Arts & Humanities Citation Index Expanded was not covered, because the authors' institutions did not have access. But in addition to the Science Citation Index Expanded (SCIE) and the Social Science Citation Index (SSCI) (which were also used in 2010), both conference proceedings citation indexes (Science and Social Sciences) were covered by the search queries. This was done in order to get a more complete coverage of the publication output of the former Yugoslav countries. Again similar to 2010, the search queries consisted of the names of cities from the former Yugoslav countries, since before 1990 all successor states belonged to SFRY. In 2010, a total of 133 city and town names were used in the search queries (including synonyms of city names). For the current study, we also used search queries that consisted of the country names (Yugoslavia and all successor states) in order to find city and town names (and synonyms), which were missing in our city search queries. In addition to that, the maximum number of 50 search arguments in WoS (still existing in 2010) is no longer limited which meant that we were able to use much longer search queries for the current study. Because of this, the new search query included 769 city and town names along with synonyms, misspellings etc. This has led to a much broader database and a better allocation of publications to their respective states, in comparison to the data used in 2010. In 2010, the data set consisted of 103 963 publications (for the years 1970 to 2007), the current study has 121 602 publications for this time period (20% more) plus 101 533 publications for the years 2008 to 2014, which brings the complete data set to a total of 223 135 publications. We rechecked whether these publications were all from the correct countries by using WoS exclude tool and removing all publications from the seven Yugoslav successor states. The remaining publications consisted of around 1% of the total data set and manual checks of these publications have shown that most of these were still relevant but wrongly indexed (for example publications from Kosovo which were attributed to Albania). This leads us to believe that our data set includes all publications from the former SFRY, which can be found in the WoS.

We analysed the data set using a proprietary bibliometry toolbox (programmed at Fraunhofer INT) and the following measures and method: (1) Absolute number of publications for each state (2) Absolute number of cooperation for each state and (3) Visualization of the Yugoslav cooperation network. In our future studies, we will add measures like Salton's measure and others.

Results

Results of the analysis are represented by four figures depicting cooperation networks between former Yugoslav republics and the province of Kosovo for the periods before the Yugoslav wars (from 1970 until 1990), during the wars (from 1991 until 1999), in the first decade after the wars (from 2000 until 2009), and in the last 5 years (from 2010 until 2014). Each republic's and the province of Kosovo's publications indexed by WoS have been represented in figures by a circle which size is proportional with the number of publications published by researchers from each respective republic. Lines between those circles represent cooperation of researchers in writing publications and line thickness is proportional with the number of collaborative publications of researchers from two republics whose circles are connected by the line. A cooperation was counted whenever more than one institution that published a paper was located on the territory of the former Yugoslavia and these institutions were not from the same republic. Cooperation between three or more republics are quite rare. These were enumerated as a set of multiple bilateral cooperation.

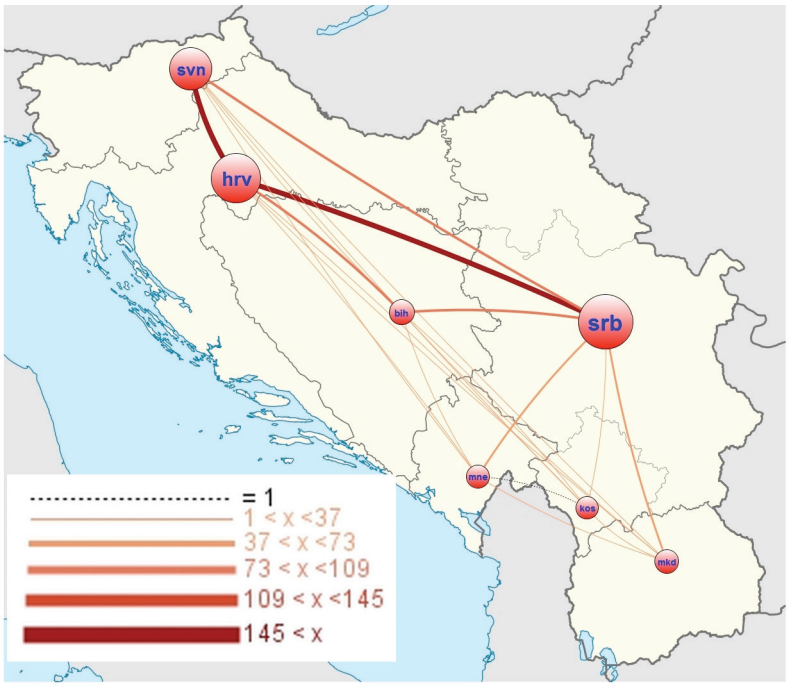


Figure 1. Visualisation of the cooperation network for 1970-1990 (before Yugoslav wars).

Figure 1 depicts the cooperation network for the period before the Yugoslav wars. Researchers from Serbia published the highest number of publications before the wars, followed by researchers from Croatia. Those two republics were the most productive republics and cooperated the most in former Yugoslavia. Slovenia, according to the productivity of its researchers and to the cooperation in this period, was in the middle between the groups of "big" republics by scientific productivity (Serbia and Croatia) and the group of "small" republics (Bosnia and Herzegovina, Macedonia, Montenegro and the province of Kosovo). Before the war, the most productive "small" republic was Bosnia and Herzegovina.

The Yugoslav wars started in 1991 and they led to a strong decrease of scientific cooperation in the republics in the 90's. Also, it affected the ratio of scientific productivity between republics during the wars. Figure 2 depicts the cooperation network for the period 1991-1999 which is the period of Yugoslav wars. Before the wars, Serbia was cooperating strongly with Croatia, Slovenia and Bosnia and Herzegovina. The cooperation triangle between Serbia, Croatia and Slovenia almost disappeared in the 90's, as well as the cooperation triangle between Serbia, Croatia and Bosnia and Herzegovina. However, scientific cooperation between Croatia and Slovenia was strengthened in this period. The reason for that is the fact that the conflict between Serbia, Croatia and Bosnia and Herzegovina during the wars was much stronger than the conflict between Croatia and Slovenia. Also, effects of the wars were much less on Slovenian economy than on the economies of other republics. War in Slovenia ended after ten days in 1991. Also, Macedonia remained at peace throughout the Yugoslav wars and declared its independence in September of 1991. Thus, the ratio of scientific productivity of Slovenian and Macedonian researchers in comparison to the other republics researchers had been changed in favour of Slovenia and Macedonia. In this period and in the followings periods Slovenia became a member of the group of "big" republics.

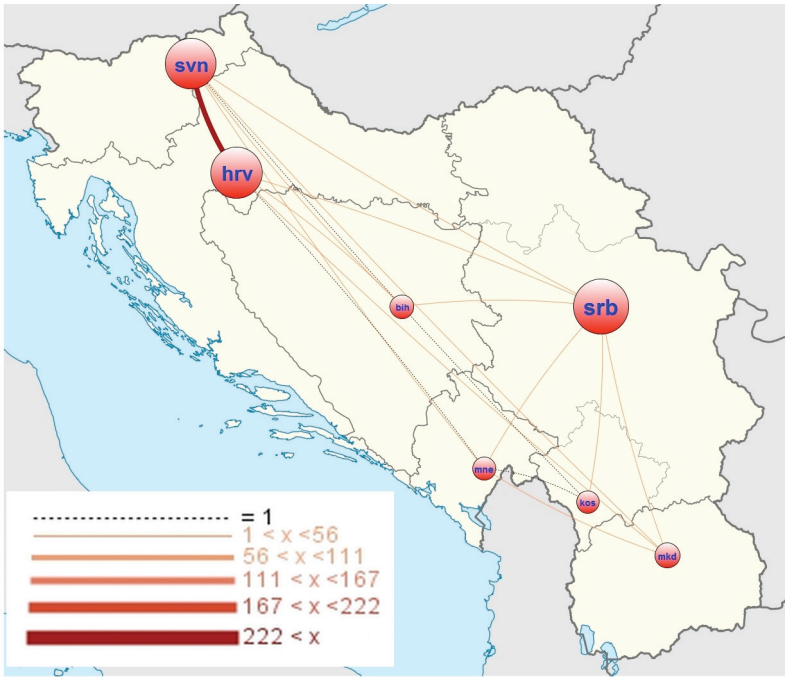


Figure 2. Visualisation of the cooperation network for 1991-1999 (during Yugoslav wars).

Figure 3 depicts the cooperation network for the period 2000-2009 which is the first decade after the Yugoslav wars. Scientific cooperation in this period between Serbia and Slovenia was strengthened again. The cooperation triangle between Serbia, Croatia and Slovenia was not as strong as before the wars (taking into account that the overall publication output increased), but it seems as if this cooperation triangle was resurfacing again.

Figure 4 depicts the cooperation network for the period 2009-2014. In this period Serbia has returned to having the most publications as before the Yugoslav wars. Reasons for this include introduction of a new rulebook for evaluation prescribed by the Ministry of Education, Science and Technological Development of the Republic of Serbia in 2008. That rulebook requires researchers must have articles published in journals in the Web of Science database for the promotion to scientific positions. In addition, the increase in the number of publications was influenced by the fact that several journals based in Serbia have, in recent years, started to be indexed by Web of Science: e. g. Vojnosanitetski Pregled, Archives of

Biological Sciences, Srpski Arhiv Za Celokupno Lekarstvo, Journal of the Serbian Chemical Society, etc. Those journals published a considerable number of articles written by Serbian researchers in the period 2010-2014 (Ivanović and Ho, 2014). The strengthening of the cooperation triangle between Serbia, Croatia and Slovenia started in the period 2000-2009 continues in the last five years. We conclude that this triangle is fully recovered.

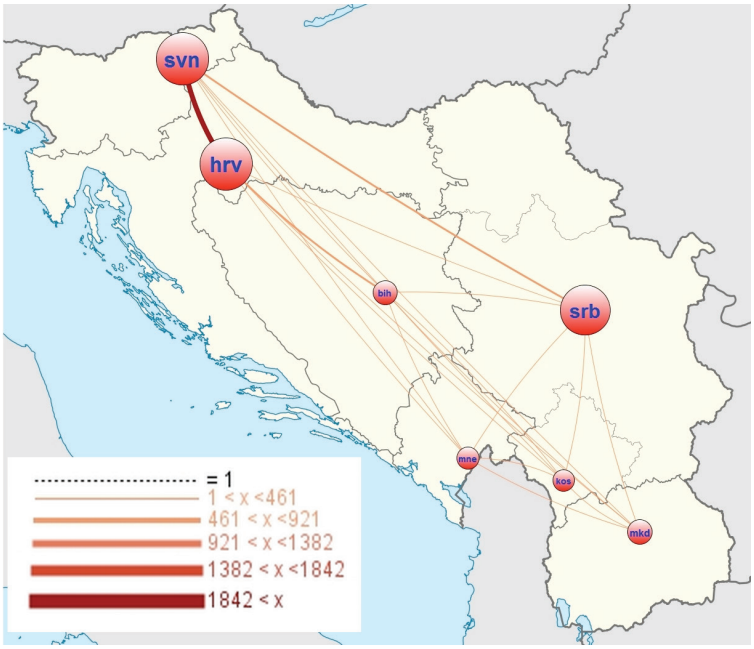


Figure 3. Visualisation of the cooperation network for 2000-2009 (1st decade after Yugoslav wars).

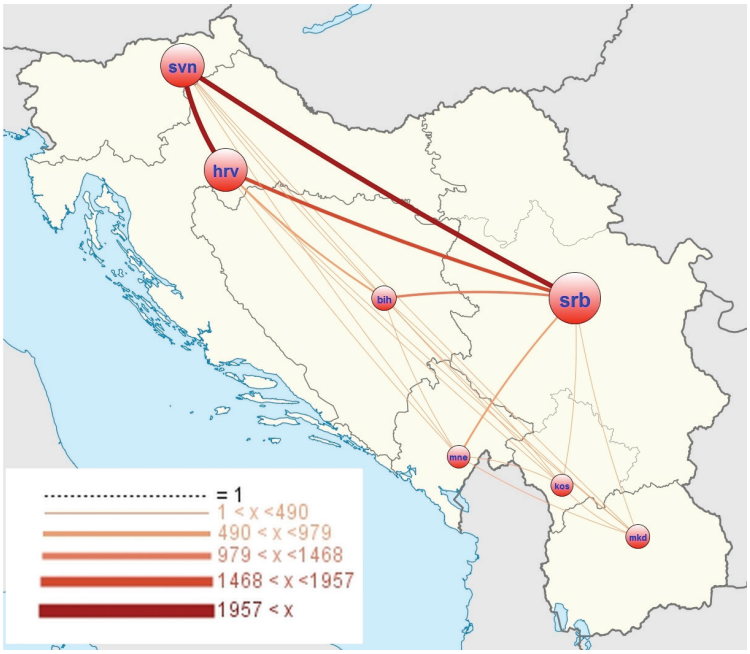


Figure 4. Visualisation of the cooperation network for 2010-2014.

Conclusion

The analysis of scientific-research outputs of the republics of former Yugoslavia for the period 1970-2014 has been presented in this paper. It reveals that civil Yugoslav wars affected the republics' productivities and scientific cooperation in different ways. The most

affected republics by wars and social crisis were Serbia and Bosnia and Herzegovina, while the least affected republics were Slovenia and Macedonia. However, it seems that in the last five years productivity and scientific cooperation look similar as before the Yugoslav wars. This result strengthens the results from the 2010 study. It would seem that old cooperation networks, which were disrupted during the Yugoslav wars, are in place again. However, our data cannot answer the question whether these are the same networks as before (i. e. the same researchers and/or institutions that are cooperating again) or whether new ones have taken the place of the old ones.

The presented results are the first part of our research. We are going to extend our research with following measures and methods: relative number of publications for each state and normalized cooperation score $R_i^{(cs)}$ (as described in Jovanović et al. (2010). Also, we are going to analyse the distribution of collaborative articles per the biggest Universities based in these states.

References

Aksnes, D. W. (2003). Characteristics of highly cited papers. *Research Evaluation*, *12*(3), 159-170.

Baltussen, A., & Kindler, C. H. (2004). Citation classics in anesthetic journals.*Anesthesia & Analgesia*, *98*(2), 443-451.

de Bruin, R. E., Braam, R. R., & Moed, H. F. (1991). Bibliometric lines in the sand. *Nature*, *349*, 559-562.

Đukić, V., Udiljak, N., Bartolić, N., Vargović, M., Kuduz, R., Boban, N., Pećina, M. & Polašek, O. (2011). Surgical Scientific Publication and the 1991-1995 War in Croatia. *Collegium Antropologicum,* 35(2), 409-412.

Figg, W. D., Dunn, L., Liewehr, D. J., Steinberg, S. M., Thurman, P. W., Barrett, J. C., & Birkinshaw, J. (2006). Scientific collaboration results in higher citation rates of published articles. *Pharmacotherapy: The Journal of Human Pharmacology and Drug Therapy*, *26*, 759-767.

Fu, H. Z., & Ho, Y. S. (2013). Independent research of China in Science Citation Index Expanded during 1980–2011. *Journal of Informetrics*, *7*(1), 210-222.

Furlan, J. C., & Fehlings, M. G. (2006). A web-based systematic review on traumatic spinal cord injury comparing the" citation classics" with the consumers' perspectives. *Journal of neurotrauma*, *23*(2), 156-169.

Ho, Y. S., Satoh, H., & Lin, S. Y. (2010). Japanese lung cancer research trends and performance in Science Citation Index. *Internal Medicine*, *49*(20), 2219-2228.

Igić, R. (2002). The influence of the civil war in Yugoslavia on publishing in peer-reviewed journals. *Scientometrics*, *53*(3), 447-452.

Ivanović, D., & Ho, Y. S. (2014). Independent publications from Serbia in the Science Citation Index Expanded: a bibliometric analysis. *Scientometrics*,*101*(1), 603-622.

Jovanović, M. M., John, M., & Reschke, S. (2010). Effects of civil war: scientific cooperation in the republics of the former Yugoslavia and the province of Kosovo. *Scientometrics*, *82*(3), 627-645.

Kutlača, D., Babić, D., Živković, L. & Štrbac, D. (2015). Analysis of quantitative and qualitative indicators of SEE countries scientific output. *Scientometrics,* 102, 247-265

Leta, J., & Chaimovich, H. (2002). Recognition and international collaboration: the Brazilian case. *Scientometrics*, *53*(3), 325-335.

Lewison, G., & Igic, R. (1999). Yogoslav politics,“ethnic cleansing” and co-authorship in science. *Scientometrics*, *44*(2), 183-192.

Lukenda, J. (2006). Influence of the 1991-1995 war on Croatian publications in the MEDLINE database. *Scientometrics*, *69*(1), 21-36.

Moravcsik, M. J. (1985). Applied scientometrics: an assessment methodology for developing countries. *Scientometrics*, *7*(3-6), 165-176.

Wagner, C. S., & Leydesdorff, L. (2005). Network structure, self-organization, and the growth of international collaboration in science. *Research policy*,*34*(10), 1608-1618.

The Brazilian National Impact: Movement of Journals Between Bradford Zones of Production and Consumption

Rogério Mugnaini[1] and Luciano A. Digiampietri[2]

[1] *mugnaini@usp.br*
University of São Paulo, School of Communication and Arts (ECA), Av. Prof. Lúcio Martins Rodrigues 443, 05508-020 São Paulo (Brazil)

[2] *digiampietri@usp.br*
University of São Paulo, School of Arts, Sciences and Humanities (EACH), Av. Arlindo Bettio 1000, 03828-000 São Paulo (Brazil)

Abstract
A specific aspect of the scientific communication in non-English-speaking countries is the need for insertion in the global knowledge flows since a significant part of their publications occurs in national or regional journals. This had led many countries to create alternative ways to assess national journals, allowing a more trustworthy view of the national scientific production. This study aimed to characterize the journals used in the Brazilian scientific production in Web of Science and SciELO, in order to observe the dynamics along five triennia and across the Bradford Zones for both production and consumption in the different areas. Bradford zones showed to be an interesting relative indicator, when applied to evaluative purposes. Especially the joint analysis of production and consumption dimensions can bring a more complete view of the scientific communication system, and this study showed the flows of journals through zones in both dimensions.

Conference Topic
Country-level studies

Introduction
In the last years, several efforts were undertaken by the developing countries in order to improve their position in the global scientific scenario. However, as important as (or even more important than) improve their position is to formulate and implement initiatives for improving their research system, in which the scientific communication plays important role.

A specific aspect of the scientific communication in these countries, mainly in the non-English-speaking ones, is the need for insertion in the global knowledge flows (Ponomariov & Toivanen, 2014), because a significant part of their publications occurs in national or regional journals (Mugnaini et al., 2014). The researchers from these countries, many of them involved in scientific editing, face the dilemma between maximizing efforts to publish in *mainstream journals* and improve the national journals in order to internationalize them – and its negative consequences of such a process (Rego, 2014). Both aspects are typically treated as ways to internationalize the national science, but is this enough (Buela-Casal et al., 2006)? This duality comes from the national science policy, which in one hand valorizes the journals with high Impact Factor (IF) and, on the other hand, tries to attend the clamor for recognition of the national journals (Miranda & Mugnaini, 2013).

This had led many countries to create alternative ways to assess or classify the national journals, allowing a more trustworthy view of the national scientific production, identifying the role of the national journals. In order to do this, some countries built national citations indexes: SciELO Project (Packer et al., 1998), Chinese Science Citation Database (Jim & Wang, 1999), Korea Citation Index (Kim et al., 2013), Citation database for Japanese papers (Negishi et al., 2004) and Islamic World Science Citation Center (Mehrad & Arastoopoor, 2012). Other countries considered this kind of initiative as a solution only for the Humanities and Social Sciences, and are looking for different ways to include the national journals in their scientific evaluation process: Taiwan (Chen, 2004), Spain (Piñeiro & Ricks, 2015),

Poland (Winklawska, 1996), Serbia (Šipka, 2005), among other countries from Eastern Europe (Pajić, 2014) and a project originally european – European Reference Index for the Humanities and the Social Sciences-ERIH PLUS – which currently reaches worldwide.

By the way, despite being considered, national journals are minimally punctuated in comparison to journals indexed in WoS. One of the reasons of this non-recognition is the fact that many of these journals are not peer-reviewed, and, among the ones that are, some present and endogen editorial board (Packer, 2014). These facts explain the non-inclusion of these journals in the most recognized citation databases. Consequently, the commissions of researchers that tread the paths of the national research assessment exercise have to deal with these characteristics as extra factors. On the other hand, the creation of national data sources with defined selection process can be a solution.

The limited insertion of these countries' research in mainstream science finds no echo (Tijssen et al., 2006), since it lacks potential audience (MacRoberts & MacRoberts, 1996), indispensable to a consistent citation analysis. Thus, the evaluation is based strictly on productivity indicators, which impose even bigger challenge to establishing quality criteria. Therefore it became necessary the classification of the journals. A side effect of this is the need, for these researchers who work in a research area with local/regional focus (as typically occurs in Social Sciences and Humanities), to publish a significantly higher number of papers, inflating the entire scholarly communication system (Rego, 2014).

The journals evaluation performed by CAPES in Brazil fit these aspects and have considerably different criteria among the 48 areas (Miranda & Mugnaini, 2013). The most common criteria are (sorted in a decreasing way, according with the assigned importance): citation indicators (JCR Impact Factor, Scopus/SCImago or Google Scholar H-index, SCImago Journal Ranking, or a mix of more than one); indexing in databases with explicit selection criteria (such as Web of Science, Scopus, SciELO, thematic bases - e.g. MEDLINE, or regionals – such as, Redalyc, Latindex) or without explicit selection criteria (e.g. PASCAL); journals characteristics. All the journals where Brazilian researchers published their papers during the preceding triennium are classified. Some journals can receive different classifications from different areas (e.g. Cadernos de Saúde Publica).

Considering this scenario, stands out the need to complement the range of citation indicators for journals classification, providing a consistent view to the national context. In order to fulfill this need, in this paper a nationally recognized base - whose selection process considers explicit criteria – were created aggregating the national scientific production from SciELO and WoS (including the publications bibliographic references). The papers from this base were used to evaluate the national production and the references to evaluate the consumption. The former indicates the utility of each journal for its area; the latter indicates its impact. For both, the Bradford Zones (BZs) were calculated for each area and triennium.

This study aims to characterize the journals dynamics along five triennia and across the Bradford Zones for both production and consumption in the different areas. This study also searched for specific behaviors when comparing the journals from Brazil, from Latin America, and from the rest of the world. Other aspect analyzed was the temporal relationship in the climbs for the journals that presented climbs in both: production and consumption.

Methods

We retrieved the articles of Brazilian authors from Web of Science (WoS) and SciELO databases in a fifteen years period (1998 and 2012) - five triennia that match the national assessment exercise performed by CAPES. It was called production (PROD) data set, with 395,650 articles, published in 9,092 journals. WoS journals cover 56.4% of the articles, while 12.5% came from SciELO journals, and 28.8% from journals indexed in both databases. The remainder 23% came from journals indexed in SciELO in less than a half of a triennium

period, getting "not indexed" in such triennium - likewise, some SciELO journals turned SciELO/WoS in a triennial transition. We classified the journals using the Science Watch (2014) schema that relates WoS categories to 22 Essential Science Indicators categories, to which we added the Human Sciences. SciELO journals were classified at the same way.

Respectively, de consumption (CONS) data set was formed by 10,759,279 bibliographic references of the articles. In the case of SciELO, we just added references related to journals, but WoS data include references to proceedings, and sometimes, to thesis. These citations remained in such amount once it was discarded in the normalization process (described below) that resolved 71.3% of the references (7.67 million), as presented in Table 1.

For this first approach, we decided to restrict CONS information to citations directed to those titles that belong to PROD data set. The reason was the fact that we have almost 29% of total references not normalized automatically, and that PROD journals capture 90.3% of the normalized citation amount.

Table 1. Consumption data sets and its prevalence in the whole data set.

CONS data set (filters)	Citation window	Freq.	*% of* All citations	*% of* Normalized	*% of* Citations to PROD journals from any area	*% of* Citations to PROD journals restricted to
All citations	all	10,759,279	*100.0%*			
	5 year	3,731,745	*34.7%*			
Normalized cited journal titles	all	7,666,238	*71.3%*	*100.0%*		
	5 year	2,777,013	*25.8%*	*36.2%*		
Citations to PROD journals, from any area	all	6,922,780	*64.3%*	*90.3%*	*100.0%*	
	5 year	2,655,547	*24.7%*	*34.6%*	*38.4%*	
Citations to PROD journals, restricted to its own area	all	3,748,044	*34.8%*	*48.9%*	*54.1%*	*100.0%*
	5 year	1,485,463	*13.8%*	*19.4%*	*21.5%*	*39.6%*

So we created four different CONS data sets (featured in bold in Tab. 1), resulting of crossing two dummy variables. The first one was the restriction or not of the citation window (all citations/5-year). The second concerns to the area from which the citation comes to one title. In one case we considered just the citation received from titles of the same ESI category (not too restrictive, since it aggregates lot of WoS categories). In the other case, we count the citations regardless the area. The former corresponds to 54.1% of the latter. To give an idea of our purpose on doing this, we calculated the share of citations each area receives on its own area. The first one in the list was Space Science (whose impact is the most endogenous, with 81.2%) and the last is Multidisciplinary (the least endogenous, as one can expect, with 2.3%).

The cited journal title normalization has been performed relating the ways a journal was cited by the papers' authors with a reference base which contains several variations of cited journal title for each journal obtained from different databases (ISSN, WoS, Scopus, SciELO and Lattes Platform). Thus, it was possible to identify the ISSN from the most of the cited journals. Whenever there were conflicts in this identification, i.e., the cited title could be referring to more than one journal, the year and volume of the publication was used. In order to do this, a database containing the valid years and volumes for each journal was created using information available from the citations were the normalization presented no conflict. If, even after the use of year and volume, the conflict persisted, the normalization was not performed for the respective citation.

Having the normalized data from PROD and CONS from the 9,092 journals, as well as their basic information (title, ISSN, classification area and citing and cited years) we identified BZs, with three partitions, for which of the 23 areas in each of the 5 triennia, totalizing 115 Bradford's distributions for PROD data set. In the case of CONS data sets we did the same,

but four times, resulting 460 distributions. Moreover, it was not assigned a BZ for the journals without production or consumption in a given triennium.

An initial analysis suggested some journals had to be discarded because there was not enough information to correctly identify the behavior of these journals along the triennia. It was the case of 2,376 journals that entered the PROD data set in the last two triennia (publishing less than ten papers per triennium). An opposite case consists of 39 journals that the community stopped publishing, having no publications in the last triennium. We also found 247 journals with no articles in four triennia, and no citation in four of five triennia. Without these exclusions, 6,492 journals remained in the analysis.

The dynamics of each journal across BZs in its area was assessed along the triennia. Journals without any change in the BZ along the five triennia were classified as Stable (S). The ones that climbed zones along the triennia without any fall were considered Up (U), and oppositely, journals that fell BZs across the triennia without any climb were considered Down (D). And a journal that had climbs and falls along the triennia was considered Oscillating (O).

Findings

The great amount of data demanded many cross-tabulations to define the way of treating the information of each variable. At this time, we decided not to differentiate if a journal climbed one (Z3 to Z2 or Z2 to Z1) or two (Z3 to Z1, in different triennium or in a unique double step). The same was proceeded in relation to journals that fell BZs.

As we needed to create a journal profile of change that combine both PROD and CONS, we aggregated it with the following ordered classification scheme: U, to any combination that occurred at least one Up, permitting one of them to be Stable (U-U, U-S or S-U, to both PROD and CONS, respectively); S-S, if the journal has being Stable in both dimensions; O, if it was found swinging in any of dimensions; and D, to any combination occurring a Down.

Table 2. Distribution of journals by profile of changes in Bradford zones of production and consumption, in the four CONS data sets – period 1998-2012.

Citation data sets	Journals (total)					
Publication country	U	S_S	O	D	%	Freq.
CONS, considering citations to PROD journals, from all areas						
all	***10.8%***	***76.1%***	***8.7%***	***4.4%***	***100.0%***	**6,492**
Other	*9.5%*	*77.3%*	*8.7%*	*4.4%*	*100.0%*	5,949
Latin Am. &Caribe	*2.6%*	*93.1%*	*3.4%*	*0.9%*	*100.0%*	233
Brazil	*41.0%*	*39.7%*	*11.0%*	*8.4%*	*100.0%*	310
5 year	***10.4%***	***73.4%***	***10.6%***	***5.6%***	***100.0%***	**6,410**
Other	*9.3%*	*74.5%*	*10.6%*	*5.7%*	*100.0%*	5,873
Latin Am. &Caribe	*3.1%*	*92.5%*	*3.9%*	*0.4%*	*100.0%*	228
Brazil	*38.2%*	*38.2%*	*14.9%*	*8.7%*	*100.0%*	309
CONS, considering citation to PROD journals, restricted to its own area						
all	***17.7%***	***65.0%***	***12.1%***	***5.3%***	***100.0%***	**6,430**
Other	*16.5%*	*65.8%*	*12.4%*	*5.2%*	*100.0%*	5,890
Latin Am. &Caribe	*3.9%*	*90.9%*	*3.9%*	*1.3%*	*100.0%*	232
Brazil	*50.3%*	*28.6%*	*12.3%*	*8.8%*	*100.0%*	308
5 year	***16.8%***	***60.9%***	***15.2%***	***7.0%***	***100.0%***	**6,310**
Other	*15.8%*	*61.6%*	*15.6%*	*7.0%*	*100.0%*	5,777
Latin Am. &Caribe	*4.8%*	*90.3%*	*3.5%*	*1.3%*	*100.0%*	227
Brazil	*45.8%*	*25.8%*	*17.6%*	*10.8%*	*100.0%*	306

So we first have looked to the general behavior of the journals, but focusing on the ones that improved across the triennia, at least in one of the dimensions. Tab. 2 shows that the great amount of journals (about 75%) are Stable in both dimensions, but we find 10% less journals

with this profile when we restrict the citations to the journals own area. It reveals that closing the context of citation to the specific area, we find more changes (and this tendency is even more evident in the 5-year citation window), especially for the journals that got climbed BZs.

Considering the publication country, we can realize that Brazilian journals present lesser stability, what is interesting to analyze changes, which is what we find abundantly: about 40% when considering citation from any area, and about 50% in the journals own area. Revealing the importance of studying the impact of these journals in their context.

Despite being less frequent, journals falling are more prevalent in the 5-year citation window.

All this tendencies have to be analyzed more carefully subsequently, since specific characteristics of the journals can help to understand such evidences.

Now focusing our analysis in U-U journals, it is important to mention that Clinical Medicine presents more journals (about 30), followed by Engineering (about 15), and in the opposite side is Physics (with 2). Another observation is that U-U Brazilian journals correspond to 14.5%, considering citations from all areas, and 18% in the journals own area. This is strongly different of journals out of Latin America & Caribe, whose correspondent percentage is about 3%. Among Brazilian journals, those indexed just in SciELO presents prevalence about 5% bigger than those indexed in both databases, when considering the citations in the journals own area. It reveals the growing importance of some journals in the national context, inside the area of specialty (data not shown).

Table 3. Distribution of journals U-U by triennium of first climb in Bradford zones of production and consumption, in the four CONS data sets – period 1998-2012.

Citation data sets		*% of journals: t* riennium of 1st climb in BZs (CONS)				Journals (total)	
U-U Journals		2	3	4	5	%	Freq.
CONS, considering citations to PROD journals, from all areas							
all		***11.4%***	***24.6%***	***29.8%***	***34.2%***	***100.0%***	**114**
Triennium of 1st climb in BZs (PROD)	**2**	*17.2%*	*44.8%*	*20.7%*	*17.2%*	*100.0%*	29
	3	*12.9%*	*29.0%*	*32.3%*	*25.8%*	*100.0%*	31
	4	*5.3%*	*10.5%*	*34.2%*	*50.0%*	*100.0%*	38
	5	*12.5%*	*12.5%*	*31.3%*	*43.8%*	*100.0%*	16
5 year		***14.1%***	***25.6%***	***30.1%***	***30.1%***	***100.0%***	**156**
Triennium of 1st climb in BZs (PROD)	**2**	*34.5%*	*37.9%*	*13.8%*	*13.8%*	*100.0%*	29
	3	*19.5%*	*41.5%*	*31.7%*	*7.3%*	*100.0%*	41
	4	*3.7%*	*11.1%*	*40.7%*	*44.4%*	*100.0%*	54
	5	*6.3%*	*18.8%*	*25.0%*	*50.0%*	*100.0%*	32
CONS, considering citation to PROD journals, restricted to its own area							
all		***18.5%***	***24.3%***	***27.2%***	***30.1%***	***100.0%***	**173**
Triennium of 1st climb in BZs (PROD)	**2**	*41.4%*	*31.0%*	*20.7%*	*6.9%*	*100.0%*	29
	3	*8.6%*	*51.4%*	*28.6%*	*11.4%*	*100.0%*	35
	4	*22.2%*	*20.4%*	*31.5%*	*25.9%*	*100.0%*	54
	5	*9.1%*	*7.3%*	*25.5%*	*58.2%*	*100.0%*	55
5 year		***22.9%***	***24.0%***	***29.6%***	***23.5%***	***100.0%***	**179**
Triennium of 1st climb in BZs (PROD)	**2**	*44.0%*	*32.0%*	*24.0%*	*0.0%*	*100.0%*	25
	3	*20.0%*	*48.6%*	*25.7%*	*5.7%*	*100.0%*	35
	4	*27.3%*	*20.0%*	*40.0%*	*12.7%*	*100.0%*	55
	5	*12.5%*	*10.9%*	*25.0%*	*51.6%*	*100.0%*	64

Attempting to the temporal relation between Ups in PROD and CONS BZs, we performed a bivariate analysis considering the triennium each journal had its first climb in BZs. Tab. 3 presents the distribution of journals of different triennia of CONS (columns), related to each triennium of PROD (lines). The row cells with bigger prevalence of journals are identified in grey scale The row cells with bigger prevalence of journals are identified in grey scale. In the first CONS data set, considering the first line, that respect to 29 journals that climbed BZs first time in the 2nd triennium, we see that most of the journals climbed in CONS in the 3rd,

followed by the 4th. It shows that most of them improved CONS BZs after (as to say, both of them above the principal diagonal). When we drop to the next lines the two more prevalent cells change to the diagonal and one before. The same can be observed in the second CONS data set (5-year citation window) and a little bit more concentrated in the principal diagonal when restricting the citation to the journals own area. Maybe in subsequent analysis we can verify properly if the increasing of consumption is pulling the increasing of production.

Final remarks

As we can observe in this first approach, a national system combining publications from both contexts (national and international) can be a useful tool to research evaluation. Bradford zones showed to be an interesting relative indicator, when applied to evaluative purposes. Especially the joint analysis of production and consumption dimensions can bring a more complete view of the scientific communication flow, considering the changes of journals through zones in both dimensions. National impact indicators can complement Impact Factor, in the sense it can add the local importance, as observed about SciELO journals.

Acknowledgments

We acknowledge FAPESP, that supports the *Young Investigators Awards* project, entitled *Scientific assessment in Brazil: study of scientific communication in scientific areas* (Grant number 2012/00255-6) and CNPq (Research Productivity Grant 477246/2013-3).

References

Buela-Casal, G. et al. (2006). Measuring internationality: Reflections and perspectives on academic journals. *Scientometrics*, *67*(1), 45-65.

Chen, K. (2004). The construction of the Taiwan Humanities Citation Index. *Online Information Review*, *28*(6), 410 - 419.

Jin, B. & Wang, B. (1999). Chinese Science Citation Database: its construction and application. *Scientometrics*, *45*(2), 325-332.

Kim, S. et al. (2013). Korea Citation Index and Its Macro Bibliometrics. *Asian Journal of Innovation and Policy*, *2,* 194-211.

MacRoberts, M. H. & MacRoberts, B. R. (1996). Problems of citation analysis. *Scientometrics*, *36*(3), 435-444.

Mehrad, J. & Arastoopoor, S. (2012). Islamic World Science Citation Center (ISC): Evaluating Scholarly Journals Based on Citation Analysis. *Acta Inform Med. 20*(1), 40-43.

Miranda, E. C. & Mugnaini, R. (2013). Scientific policy in Brazil: Exploratory analysis of assessment criteria. *PRO INT CONF SCI INF*, *14*, 1578-1586.

Mugnaini, R. et al. (2014). Comunicação científica no Brasil (1998-2012): indexação, crescimento, fluxo e dispersão. *Transinformação*, *26*(3), 239-252.

Negishi, M., Sun, Y., & Shigi, K. (2004). Citation database for Japanese Papers: A new bibliometric tool for Japanese academic society. *Scientometrics*, *60*(3), 333-351.

Packer, A. L. (2014). The emergence of journals of Brazil and scenarios for their future. *Educ. Pesqui*, *40*(2), 301-323.

Packer, A. L. et al. (1998). SciELO: a methodology for electronic publishing. *Ci. Inf.,* 27(2), 109-121.

Pajic, D. (2014). Globalization of the social sciences in Eastern Europe: genuine breakthrough or a slippery slope of the research evaluation practice? *Scientometrics*, *102*(3), 2131-2150.

Piñeiro, C. L. & Hicks, D. (2015). Reception of Spanish sociology by domestic and foreign audiences differs and has consequences for evaluation. *Research Evaluation*, *24*(1), 78-89.

Ponomariov, B. & Toivanen, H. (2014). Knowledge flows and bases in emerging economy innovation systems: Brazilian research 2005–2009. *Research Policy*, *43*(3), 588-596.

Rego, T. C. (2014). Productivism, research and scholarly communication: between poison and medicine. *Educação e Pesquisa*, *40*(2), 325-346.

Šipka, P. (2005). The Serbian citation index: Context and content. *PRO INT CONF SCI INF*, *10*, 710-711).

Tijssen, R. J. et al. (2006). How relevant are local scholarly journals in global science? A case study of South Africa. *Research Evaluation*, *15*(3), 163-174.

Winclawska, B. M. (1996). Polish Sociology Citation Index (Principles for creation and the first results). *Scientometrics*, *35*(3), 387-391.

Sustained Collaboration Between Researchers in Mexico and France in the Field of Chemistry

Jane M. Russell[1], Shirley Ainsworth[2] and Jesús Omar Arriaga-Pérez[2]

[1] *jrussell@unam.mx*
National Autonomous University of Mexico (UNAM), Library Science and Information Research Institute (IIBI), Ciudad Universitaria, 04510 Mexico DF (Mexico)

[2] *shirley@ibt.unam.mx oarriaga@ibt.unam.mx*
National Autonomous University of Mexico (UNAM), Institute of Biotechnology (IBT), Av. Universidad #2001, 62210 Cuernavaca, Morelos (Mexico)

Abstract
We analyse the co-authorship and publication patterns of 863 mainstream WoS papers in the field of Chemistry co-authored between Mexican and French institutions from 1984 to 2013 with the purpose of identifying and characterizing the dynamics of sustained collaborative research partnerships in the field between the two countries. From a normalized set of the most productive authors with ≥ 5 co-authorships we selected three Mexican scientists for a detailed analysis of their co-authorship network visualized using Gephi software and its development over time. The first was the most productive Mexican author from the main national university whose collaboration with France spanned the period from 1987-2012, while the second and third researchers work in provincial universities and whose collaboration with France is more recent but lasting 10 and 15 years respectively, and also continues up to the present day. Preliminary results suggest that sustained partnerships are driven by a strong central bond between the Mexican researcher and their foreign partner. In the first two cases, the bond is with directly with a French scientist but in the third, is stronger with an Italian rather than with the French counterpart.

Conference Topic
Country level studies

Introduction
A recent paper examining the main research thrusts and future challenges facing research into scientific collaboration mentions the need to characterize the factors underpinning successful collaborations and to ascertain how collaboration can benefit scientific development in the less developed countries (González Alcaide & Gómez Ferri, 2014). International collaboration is known to be especially important for countries whose scientific infrastructure and capacity can benefit from forging alliances with researchers from institutions abroad. Colombian researchers for instance were found to increase team output by almost 40% by co-authoring with overseas partners (Ordóñez-Matamoros, Cozzens & García, 2010).

We know little about the duration of international research collaboration between individual researchers in terms of the number and timeline of co-authored papers. Two decades ago a study looked at the production and duration of collaboration between researchers from institutions in Mexico and France in all scientific areas (Narvaez-Berthelemot & Russell, 1996). Chemistry was the subject of the greatest number of bilateral publications as well as having the highest continuity index defined as the number of articles (>2) in a given period, in this case 1980-1989, that were co-authored by the same groups. More recently an analysis of co-publications between the two countries from 1984 to 2010, showed that Chemistry gradually lost ground with respect to other disciplines notably Physics, even though the number of papers increased with time (Ainsworth et al., 2014).

The present research in progress sets out to characterize the publication dynamics of sustained collaborative research partnerships between Mexico and France in Chemistry in the period 1984-2013. We take as our starting point, the most productive authors in papers with at least

one author from both Mexico and France. Considering that interpersonal links are the key drivers of collaboration (Gaillard et al., 2013) we are also interested in analysing the relationship between co-authors and tracing the development of their networks over time. Another aspect of the collaboration we consider is the level of importance of the relationship with Mexico in the case of the French scientists or France for the Mexicans, for the total body of work of the key players during the same period and who might be the senior partner in the bilateral relationship. We adopt two approaches when analysing our publication and co-authorships data based on the following assumptions: 1. Sustained collaboration is characterized by a central relationship established between one Mexican and one French scientist. 2. Sustained collaboration between the two countries is characterized by a series of relationships forged with different French scientists and institutions.

Data source and methods

Data source was the Web of Science searching France and Mexico in the country field, covering the period 1984-2013, in the discipline of Chemistry. WoS journal subject categories were adapted to the RFCD classification scheme for the assignment of the discipline (Butler, Henadeera y Biglia, 2006). Records were downloaded to a local MySQL database. Author names with ≥ 5 co-authorships were normalized and assigned (often several) Scopus author ids and affiliations, given that author identification in WoS proved less than adequate for our purpose. Case studies were selected from the group of the most prolific Mexican authors with bilateral France-Mexico collaboration. For this preliminary presentation of results we have selected three case studies based on our initial analysis of their collaboration dynamics. These include the most prolific Mexican researcher and two other productive researchers from established groups with substantial French collaboration from two provincial state universities, namely Cecilio Álvarez y Toledano from the Institute of Chemistry at the big national Mexican university, Universidad Nacional Autónoma de México (UNAM), Ricardo Navarro-Mendoza from the Universidad de Guanajuato (UG) and Claudio Marcelo Zicovich-Wilson from the Universidad Autónoma del Estado de Morelos (UAEM).

The interactive visualization open source software Gephi was used to select and represent these collaborations and to show sub-networks within clusters. Co-authors involved in each of the papers were examined to characterize the temporal collaboration, and separately the normalized author information from Scopus was used to represent the importance of the Mexico-France collaboration in the main authors' output. The corresponding author of each paper was also identified.

Overall panorama of Mexico-France co-authorship in Chemistry

The number of co-authored papers in Chemistry between Mexico and France showed a steady rise from a mere two in 1984 to 54 in 2013 (Figure 1). Social network graphs (not shown here) show an increasing dense and complex series of relationships when comparing the first 15 years (1984-1993) with the second period (1994-2013).

Publication dynamics of sustained partnerships

Figure 2, divided into three decades, shows the dense network of co-authorships of Cecilio Álvarez y Toledano with French institutions during our period of study. The strongest link is with Henri Rudler of the Université Pierre et Marie Curie, Institut Parisien de Chimie Moléculaire starting in 1987, and to a lesser extent with Andrée Parlier of the same laboratory except during the middle period 1994-2003. Rubén Alfredo Toscano works in the same institute as Cecilio Álvarez y Toledano as a highly specialized technician and is a regular co-author. Of the 29 papers of Álvarez y Toledano in co-authorship with a French institution, 23 were published in co-authorship with Rudler. There was a notable pause in their collaboration

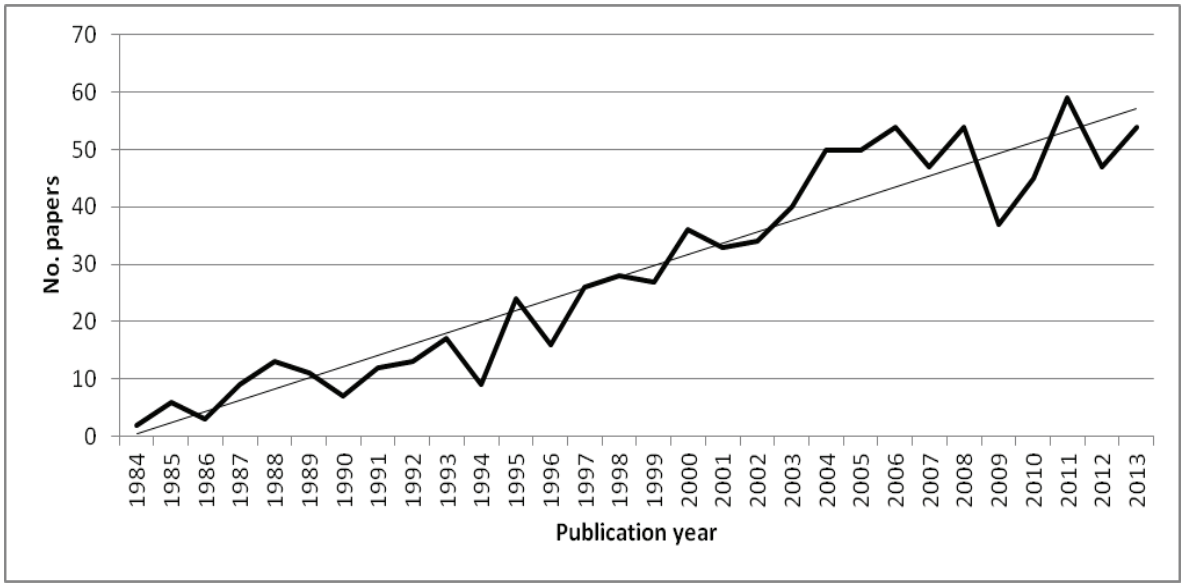


Figure 1. Papers in collaboration between Mexico and France in Chemistry 1984-2013.

from 1996 to 2004 when Álvarez y Toledano co-authored two papers with two other French authors, Henri Arzoumanian, Aix-Marseille Université, and Bruno Donnadieu now of the newly formed Université de Montpellier but at the time of the Universite Montpellier 2, respectively and involving a different set of co-authors. Nonetheless, Andrée Parlier and Henri Rudler continued their collaboration without Álvarez and Toledano during this period, together with Jacqueline Vaissermann, also from the same laboratory.

During the first two periods four clusters of co-authors are apparent, while in the most recent period 2004-2013, co-authorships are concentrated in two with Rudler and Parlier at the centre, respectively. A strong central bond with Henri Rudler is evident in the collaboration of Álvarez y Toledano over the whole period suggesting that this bilateral partnership is the motor driving this example of sustained co-authorship between Mexico and France.

Data taken from Scopus using the author id field for papers co-authored by Rudler and Álvarez-Toledano in Chemistry show Rudler to be senior (corresponding) author in 11 of these 29 papers as compared to 6 in the case of Álvarez-Toledano, which would seem to show that Rudler is the senior partner in this collaboration. The issues of authorship order are discipline-specific, but in many scientific areas it is accepted that the principal investigator is named as the corresponding author (Frandsen & Nicolaisen, 2010). These 29 papers represent 26% of all Rudler's papers as represented in Scopus, compared to 20% of those of Álvarez-Toledano suggesting that the bilateral partnership is of significance for the output in Chemistry for both researchers.

The network of collaboration with French institutions starting in 1998 around Ricardo Navarro Mendoza from the Universidad de Guanajuato appear in Figure 3 with strong links to Eric Guibal from the École des Mines d'Alès. Fourteen of the 15 papers published from 1998-2012 appear with both authors. Imelda Saucedo Medina, also from the Universidad de Guanajuato, is a co-author in 11 of these papers. In one article at the beginning of the period in 1998, there is a collaboration with other French authors, Denise Bauer and Gérard Cote, both from the École Nationale Supérieure de Chimie de Paris, and in two articles, 2000 and 2001 with Thierry Vincent from École des Mines d'Alès.

This suggests a consolidated partnership, though perhaps also an unequal one. Scopus data for papers co-authored by author Ricardo Navarro Mendoza and Eric Guibal in Chemistry show Navarro-Mendoza to be corresponding author in 8 of the 11 instances, compared to 3 for

Guibal. This would suggest that in this case the Mexican is the senior partner. These 11 papers represent 33% of all Navarro Mendoza's papers in Scopus, but only 7% of Guibal's.

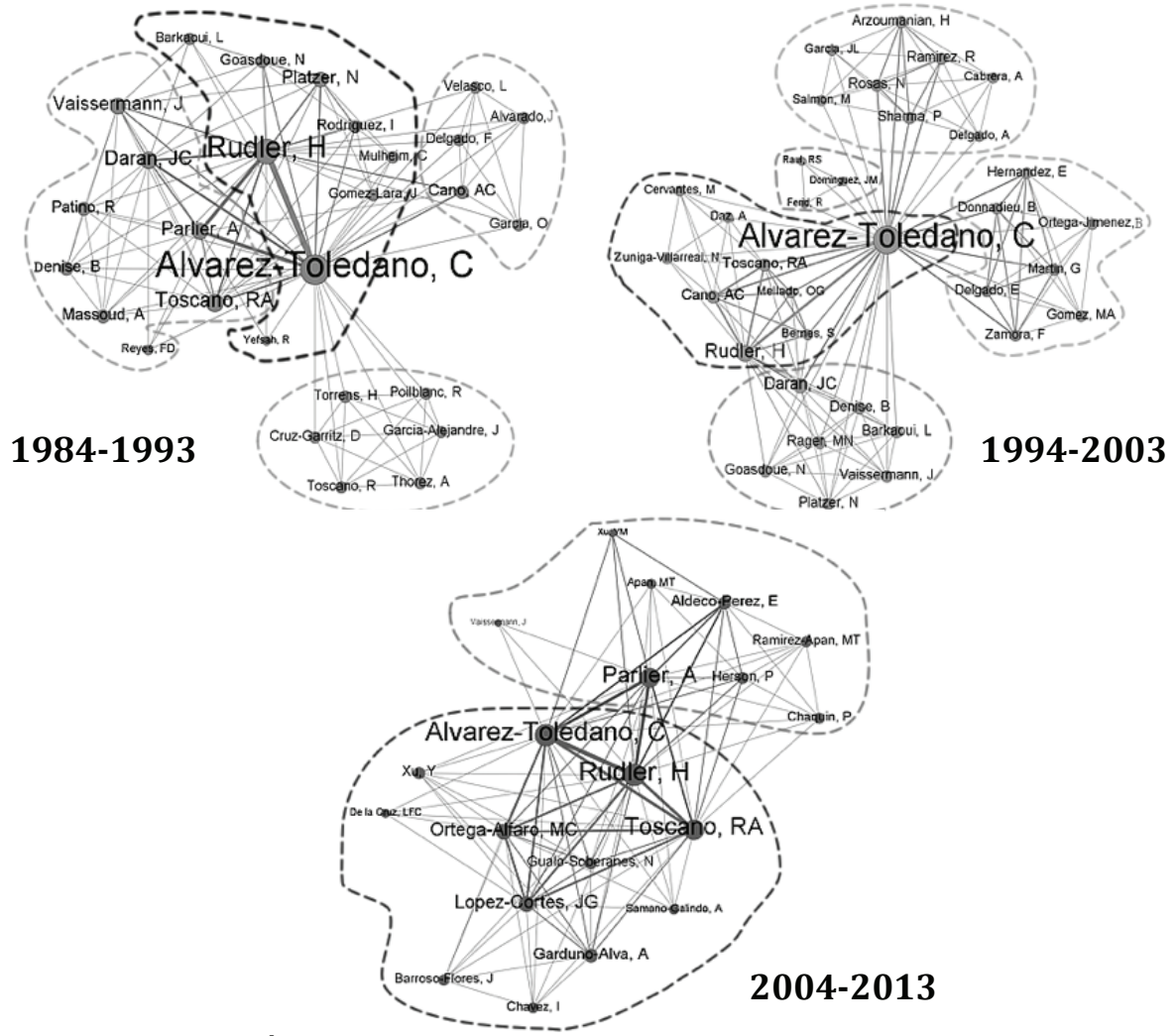


Figure 2. Álvarez y Toledano: Network of ≥ 3 co-authorships 1984-2013.

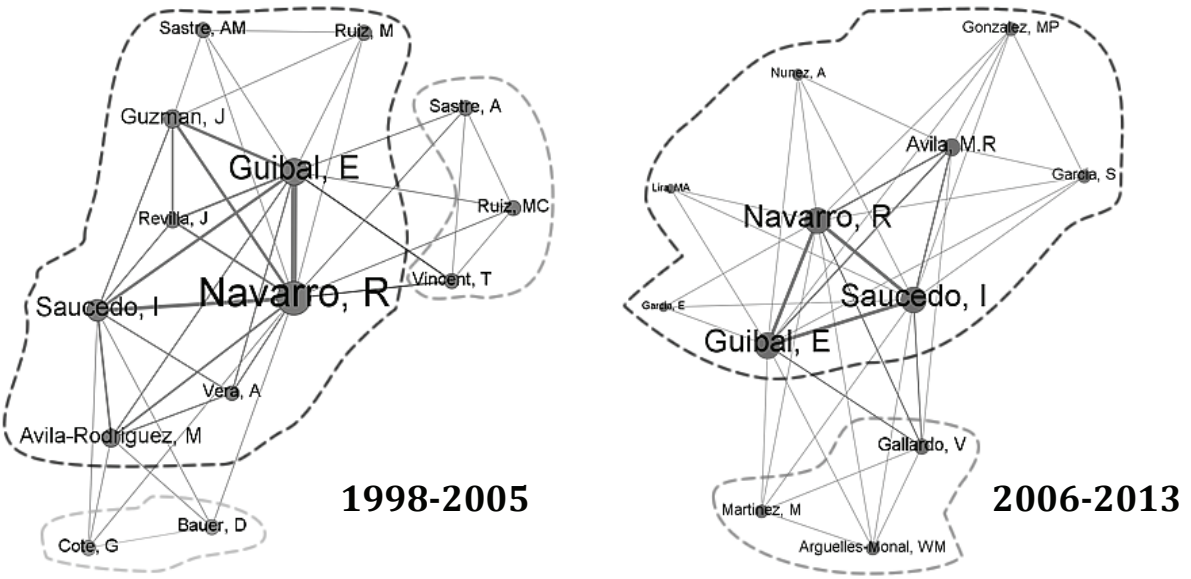


Figure 3. Navarro Mendoza: Network of ≥ 3 co-authorships 1998-2013.

The co-authorship of Claudio Marcelo Zicovich Wilson from the Universidad Autónoma del Estado de Morelos with researchers from France that began in 2004, is reflected in Figure 4, as is also the importance of a group of Italian authors for this collaboration. Roberto Dovesi from the Universita degli Studi di Torino appears as co-author in 13 of the 16 papers of Zicovich Wilson where there are also authors from French institutions in the period 2004-2013. Other researchers from the same Italian institution such as Roberto Orlando (6 papers), Piero Ugliengo (4 papers) Loredana Valenzano (3 papers 2006-2008) and Raffaella Demichelis (also 3) appear together with Dovesi, the latter co-author during 2010-2011. The predominant French author is Fabien de Pascale, at the time of Université Henri Poincaré - Nancy I, who is a co-author in 8 of the 16 papers during 2004-2010, Yves Noël, CNRS Institut des Sciences de la Terre de Paris with 5 papers 2007 then 2010-2012, together with Michel Rérat, Université de Pau et des Pays de L'Adour form a separate French collaboration, albeit together with Roberto Dovesi. The central role of Roberto Dovesi in the Mexico-France collaboration seems evident from the data taken from WoS. Data from Scopus for papers in Chemistry co-authored by Zicovich Wilson and Pascale reveal that the Mexican is corresponding author in only one of these, and Pascale not in any of them. (Pascale appears as first author in three of them.) The role of Roberto Dovesi in this collaboration seems to be confirmed in that he is corresponding author in 6 of these 10 papers. These papers correspond to 9% of all Zicovich Wilson's papers, 40% of Pascale's but only 4% of those of Dovesi. These data imply that Pascale is the junior partner here.

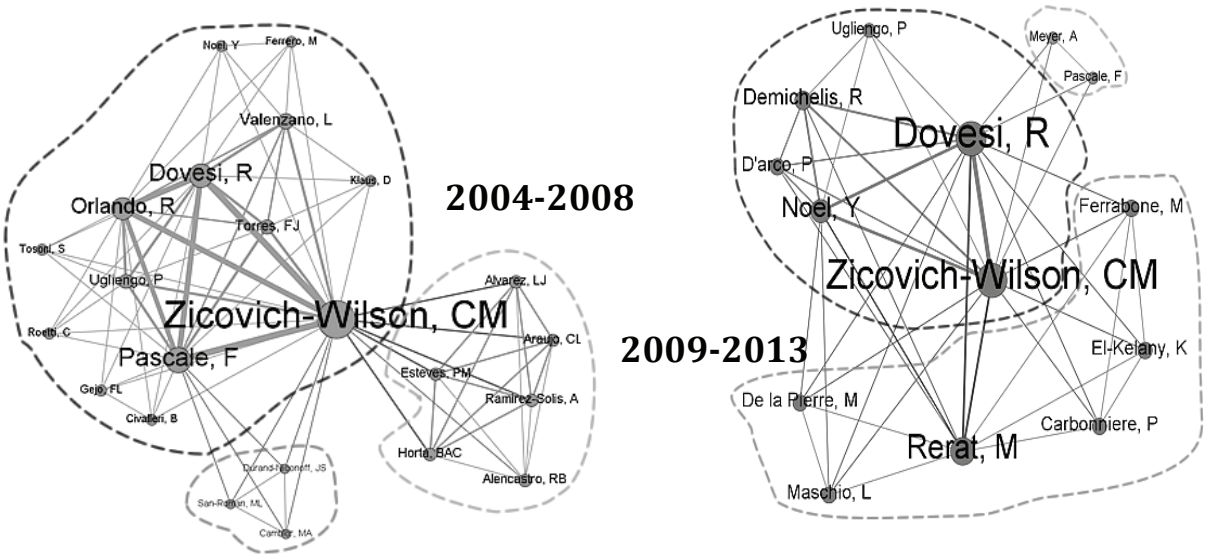


Figure 4. Zicovich Wilson: Network of ≥ 3 co-authorships 2004-2013

Preliminary Conclusions

Our detailed analyses of the co-authorship networks of three Mexican scientists, one from the large national university located in Mexico City where the national scientific research effort is centred and two from provincial universities, with ≥ 5 co-authorships with France in mainstream Chemistry journals during our period of study, lend support to our initial assumption that sustained collaboration is characterized by a central relationship established between two individual scientists but not necessarily directly between a Mexican and a French scientist. In the first two cases the bond is with a French scientist but in the third, is stronger with an Italian rather than with the French co-author. These central relationships are

strengthened and supported by frequent co-authorship from both Mexican and French groups in the first two cases and in the third, by the Italian group. A substantial number of one-time co-authors was evident in all three cases. We found differences with respect to the importance of the bilateral collaboration for the Mexican and French authors and with respect to which of the two could be considered the senior author. These preliminary conclusions will be tested by analysing further case studies of sustained partnerships between Mexican and French chemists.

References

Ainsworth, S., Russell, J.M. Narvaez-Berthelemot, N. & Arriaga-Pérez, J.O. (2014). Mapeo de la colaboración en ciencia y tecnología entre México y Francia a través de un análisis de co-publicaciones 1984-2010. In D. Villavicencio & M. Kleiche (Coords.), *Cooperación, Colaboración Científica y Movilidad Internacional en América Latina*. (pp.49-74). Buenos Aires: Consejo Latinoamericano de Ciencias Sociales-CLASCO. Retrieved January 3, 2015 from: http://www.clacso.org.ar/libreria-latinoamericana/ libro_detalle.php?orden=&id_libro=916& pagenum_rs_libros=0&totalrows_rs_libros=885

Butler, L., Henadeera, K. & Biglia, B. (2006). State and Territory based assessment of Australian research. Technical paper. Retrieved January 8, 2015 from http://www.pc.gov.au/inquiries/completed/science/technicalpaper1

Frandsen, T.F. & Nicolaisen, J, (2010). What is in a name? Credit assignment practices in different disciplines. *Journal of Informetrics*, 4, 608-617.

Gaillard, A.M., Gaillard, J., Russell, J.M., Galina, C.S., Canesse, A.A., Pellegrini, P., Ugartemendra, V. & Cárdenas, P. (2013). Drivers and outcomes of S&T international collaboration activities. A case study of biologists from Argentina, Chile, Costa Rica, Mexico and Uruguay. In J. Gaillard & R. Arvanitis (Eds.), *Research Collaborations between Europe and Latin America. Mapping and Understanding Partnership.* (pp. 157-191). Paris: Editions des Archives Contemporaines,

González Alcaide, G. & Gómez Ferri, J. (2014). La colaboración científica: principales líneas de investigación y retos de futuro. *Revista Española de Documentación Científica*, 37, 1-15.

Narvaez-Berthelemot, N. & Russell, J.M. (1996). La continuite dans la colaboration scientifique internationale: Le cas de la France et du Mexique. In R. Arvanitis & J. Gaillard (Eds.), *Memoires du Colloque ORSTOM UNESCO sur "les Sciences hors d'Occident au 20o siècle"*, Vol. 7 Coopérations Scientifiques Internationales (pp.39-52). Paris: ORSTOM.

Ordóñez-Matamoros, H.G., Cozzens, S.E. & García, M. (2010). International co-authorship and research team performance in Colombia. *Review of Research Policy*, 27,415-431.

Innovation and Economic Growth: Delineating the Impact of Large and Small Innovators in European Manufacturing

Jan-Bart Vervenne and Bart Van Looy

{jan-bart.vervenne, bart.vanlooy}@kuleuven.be
KU Leuven, Faculty of Business and Economics, Department of Managerial Economics, Strategy and Innovation, Naamsestraat 69, B-3000 Leuven (Belgium)

Abstract
In the course of the past decades, the link between innovation and economic growth has become a well-established one in the economic literature. In the current study an attempt has been provided to complement this line of research with an assessment of the wealth implications of the 'entrepreneurialisation' of innovation systems. Relying on a 9 year panel of post-millennial observations for 22 European countries and using stock based patent indicators, it was found that on top of the positive productivity impact of innovative activity growth, a premium effect can be observed when the stake of small firms in it increased at the same time. These findings can be interpreted as confirming Baumol's (2004) assignment of different roles to large and small firms in innovation systems: the former as provider of the technological breakthrough that the latter improves in a range of incremental steps. The entrepreneurialisation of manufacturing as a whole, measured by the stakes of small businesses in employment, yields a productivity discount: outside of innovative activities, economies of scale outweigh co-occurring diseconomies of scale. Distinct country groups in different stages of economic development form the main drivers of both entrepreneurialisation effects: a core of North-Western European countries that has attained the innovation-driven stage against a periphery of Southern and Eastern European countries around them that have not transcended the more preliminary efficiency-driven stage. Further rationales explaining the additional explanatory power of entrepreneurial innovation were found in the weakening of the link between innovation measured by patents and added value in large firms.

Conference Topic
Country-level studies; Patent analysis

Introduction
Substantial agreement exists among economists and policymakers that technological innovation is a key driver of sustainable economic growth. Technological innovation implies the implementation of inventions in the production of final goods or services and as such yields productivity gains for the innovating economy. Using knowledge capital to transform existing knowledge into such inventions, the amount of research and development (R&D) efforts is an important determinant of the pace of technological innovation.

Endogenous growth scholars have shown that technological innovation is an endogenous component of the process of long-run economic growth, both theoretically (Romer, 1986) as well as empirically (Nadiri, 1993). As opposed to their neoclassical counterparts (Solow, 1956), they postulate that technological innovation is an inherent component of the growth process: profit-maximising firms purposely allocate resources towards R&D in the presence of sufficient perspectives suggesting that they will be capable to appropriate the gains from it.

The analysis in this paper contributes to the mentioned line of research by complementing the measurement of overall technological innovation effects using patent statistics with an additional, patent-based indicator capturing the footprint of small, more entrepreneurial firms in the countries' stock of knowledge capital.[1] Further explanation for the rationale triggering

[1] Note that throughout this excerpt alternately we describe the firms of our interest as entrepreneurial or small. As Wennekers and Thurik (1999) argue, smallness and entrepreneurship can only be synonymous when management and ownership are not distinct. Subsidiaries of large business groups can qualify as small as well when shareholder information is not taken into account. This remark is of concern to us given the definition of small firms we will use in the empirical part (cf. below). However, given that small firms pertaining to larger

our interest to differentiate between innovation induced by small and large firms follows next. Subsequently methodology and results are reported, followed by some concluding notes. The focus on Europe in this study is justified among others by referring to the entrepreneurial innovation deficit Europe faces in comparison with the US (Veugelers, 2009).

Delineating the entrepreneurial contributions to innovation

The rationale to differentiate between incumbent and entrepreneurial innovation draws extensively from research on entrepreneurial innovation by Audretsch (2001), Baumol (2004) and Veugelers (2009). Whereas Schumpeter in 1942 predicted the gradual replacement of the entrepreneurial inventor - naturally associated with the small start-up - by routinized innovation organized by large industrials, Baumol (2004) emphasized the complementary relationship of both types of players within innovation systems. Their organizational design has induced them to specialize in different components of society's innovation process. Over the past decades revolutionary breakthrough inventions in the US have continued to come predominantly from small entrepreneurial enterprises whereas large industry have provided ever-increasing streams of incremental improvements to them multiplying capacity and speed and increasing reliability and user-friendliness. This is the result of the oligopolistic competition this relatively limited amount of very large firms, particularly in high-tech industries, engage in. It forces them to keep innovating in order to survive, but in a very risk-free and thus path-dependent way, avoiding the risks of the unknown that the revolutionary breakthrough entails. As such, inert incumbents leave plenty of room to explore for the enterprising entrepreneur. Unaffected by concerns relating to existing products and markets, the latter can pick up the ideas the former would deem too risky (Audretsch, 2001; Baumol, 2004). The other way around, incumbents are more suited to follow-up and improve those breakthrough innovations in more mature stages of the technology life-cycle (Baumol 2004).

Plugging the level of 'entrepreneurialisation' of innovation into a growth model

Methodology

The neo-classical growth model (Wong et al., 2005) we use to test a number of research questions distilled from the context described above is based on an augmented Cobb-Douglas production function:

$$Y = A^{O} K^{\propto} L^{\beta}$$

Where Y = output, A^{O}= total factor productivity, K = stock of physical capital and L = labor employed. Assuming constant returns to scale, $\alpha + \beta = 1$, both sides of the equation are then divided by labour. Taking natural logs the resulting model to estimate economic productivity per employee goes as follows:

$$\ln\left(\frac{Y}{L}\right) = lnA^{o} + \propto ln\left(\frac{K}{L}\right)$$

Following the approach by Wong et al. (2005), we assume that the stock of knowledge capital is the main determinant of total factor productivity, A^{O}. The stock of knowledge capital is captured using technological innovation statistics, among which patent based-indicators comprise one of the best proxies. More specifically, the level of innovation (*INNO)* is measured using stocks of patent applications depreciating at a rate of 20% per year as the

conglomerates in the countries of our sample never comprise a majority, on average our population of small firms can be described as 'more entrepreneurial'.

effects of investment in innovation transcend the short run.[2] The technological innovation variable was normalized by employment to capture its intensity and limit the effects of country size as much as possible. As suggested in the previous section, as factor of total productivity the general intensity of technological innovation is complemented by a patent-based indicator, measuring the degree of small firm engagement in innovative activity, and an equivalent employment-based indicator to control for overall small firm activity. The latter to make sure increased innovative activity of small firms is not simply capturing the potential productivity effects of an increase in entrepreneurial activity in general.

Determining the degree to which national innovation systems have ran on entrepreneurial initiative was based on the assignment of patents to small and large firms using the methodology presented in Eurostat (2014).[3] Due to shortcomings in the matching methodology and data gaps in the financial database - among others the result of country-specific disclosure exemptions rewarded to certain company types - only for approximately 62% of the corporate applicants in Europe firm size could be determined. We assume however that these country-level constraints equally hold for all years of the sample and as such are coped with by estimating coefficients using country fixed effects (cf. infra).

The effects of entrepreneurial and incumbent engagement in innovation could not just be measured by plugging raw stocks of their respective patent applications into the equation: R&D clustering dynamics within countries result in a high correlation – more than 0.97 even when removing country effects – with the annual innovative activity deployed by the national innovation system as a whole, that is already captured in the core variable measuring technological innovation. Given our main interest towards the benefits of entrepreneurial innovation and to avoid multicollinearity, the degree of 'entrepreneurialisation' of corporate technological innovation (*ENTR_INNO*) was measured by computing the share of small firms in the stock of patents assigned to firms with identified size.

The within variance of this share value captures to what extent small firms have shown relative over- or underactivity in R&D in comparison with their large counterparts. Given the large level of correlation among the small firm, large firm and overall patent stocks it is safe to assume that entrepreneurial and incumbent innovation do not have an opposite effect on economic productivity which would hamper a straightforward interpretation of *ENTR_INNO*. At most one of them can have a relatively larger impact on productivity. In line with the rationale elaborated above we expect that to be the small innovators. The result of that should

[2] All patent statistics were extracted from EPO's Worldwide Patent Statistical Database 'PATSTAT' (Autumn version 2014). In general we relied on EPO patent applications, including granted and non-granted patents, with the idea that counting both yields a relatively more input-oriented measure capturing the level of R&D spending than if one would stick to grants only (Ernst, 2003). Depreciation of the patent stock at a rate of 20% per year is based on the perpetual inventory method described in Ulku (2004). The patent stock variable incorporates annual EPO patent counts from 1970 onwards. The restriction of our attention to EPO patents can be easily justified given the geographical reach of our dataset and their costliness, which is a direct result of their supra-national character. Being that expensive, especially for more financially constrained SMEs, counts of them at the macro-level bear the potential to be good signals of R&D input & output levels per country over time.

[3] The lack of dynamic shareholder data in BvD's Amadeus (a database gathering annual account information) withheld us from determining firm size at the business group level. In contrast with the matching exercise presented in Eurostat (2014), firm size was determined dynamically by linking patents to financial information from the financial years that corresponded with the patent application filing year. In addition financial account data from Amadeus 2012 was enriched with equivalent information from earlier versions (2004 and 2007) to dispose of financial information in the earliest years of the matched sample (1999-2011) and to account for the BvD rule to discard companies not filing accounts for 5 years in a row. Firm size – or rather entity size – classification for patenting companies from 1999 onwards was based on the European Commission SME definition (2005): enterprises that employ fewer than 250 employees and which have an annual turnover not exceeding 50 million euro, and/or an annual balance sheet total not exceeding 43 million euro.

be *ENTR_INNO* exerting a positive effect on productivity, which would imply the existence of a productivity premium to an increased entrepreneurial stake in corporate innovation.

Given that the large majority of patents in Europe can be assigned to the manufacturing industry (Fraunhofer, 2003), downloads of observations for the non-patent based variables of country *c* in year *t* were restricted to that sector. Indicators for value added at factor cost (*VAFC*), the number of persons employees (*NPE*), gross investment in tangible goods (*GITG*) and the share of small firms in corporate employment (*ENTR_EMP*) were extracted from the Eurostat website.[4] [5] Furthermore, a quadratic year trend is included to capture time effects.[6] Conform previous research all R&D related indicators are lagged since it is assumed that the effects of R&D on economic performance take a couple of years to surface. In line with Ulku (2004) and given the limited time-series at our disposal we opted for a 2-year time lag. Following an equivalent rationale, the physical investment and share of entrepreneurial employment variables were also lagged by 1 year.

The resulting equation to be estimated using panel data techniques is:

$$lnVAFC/NPE_{ct} = \propto + \beta_1 lnGITG/NPE_{c,t-1} + \beta_3 INNO/NPE_{c,t-2} + \beta_4 ENTR\_INNO_{c,t-2} + \beta_5 ENTR\_EMP_{c,t-1} + year + year^2 + u_c + \varepsilon_{ct}$$

Results

Coefficients are estimated using fixed effects OLS.[7] Table 1 reports the estimation results, including robust standard errors, for the overall set of European countries (panel 1: ALL) and split sets of countries that lead (panel 2: LEADERS) or lag behind (panel 3: LAGGARDS) in terms of innovation according to the European Commission's (EC) Innovation Union Scoreboard (2015). The left hand of each panel contains estimates for the basic model as expressed in the equation above. The right hand side in addition reports an additional interaction effect between the technological innovation intensity and its degree of 'entrepreneurialisation'.

Conclusion and directions for future research

Apart from confirming previous findings regarding the positive impact of technological innovation on economic output, overall results (ALL) reveal that there is an additional productivity premium to a larger share of entrepreneurial engagement in the development of new, patented technology. The entrepreneurialisation of employment on the other hand, a broader measure of corporate activity, appears to be negatively associated with productivity.

---

[4] The resulting set of 22 countries consists of: Austria, Belgium, Germany, Denmark, Finland, France, United Kingdom, the Netherlands, Norway, Slovenia, Sweden (LEADERS), Czech Republic, Cyprus, Estonia, Greece, Hungary, Italy, Latvia, Poland, Portugal, Slovakia and Spain (LAGGARDS). Other European countries were discarded for multiple reasons: a lack of employment, investment or gross added value statistics available to the public or a too low rate of patenting companies matched to companies in the financial database, as such, hampering a representative image of the distribution of patents between incumbents and small businesses. Unusual annual productivity growth induced by preferential tax regimes for foreign firms, inciting those to shift profits to local subsidiaries, resulted in elimination of Ireland and Luxemburg from the sample as well.

[5] All currency-based series – expressed in Euro – were deflated using per country GDP price deflators (World Bank WDI website). Due to the lack of availability of stock variables capturing the total amount of outstanding fixed capital, in line with Ulku (2004) we used the flow variant.

[6] Preferably time dummies are included but using a functional form, in this case a quadratic trend allowing for one up and one down trend, can be an alternative in order to preserve degrees of freedom. Results turned out to be largely consistent for trend- and dummy-based models.

[7] Correlations among demeaned variables suggest that multicollinearity is not an issue for within-transformed variables.

Table 1. OLS fixed effects regression results.

	ALL		LEADERS		LAGGARDS	
ln_GITG/NPE (1y lagged)	0.676	0.529	0.686	-0.014	0.578**	0.825***
	(1.23)	(1.11)	(1.0)	(0.03)	(2.51)	(4.81)
INNO/NPE (2y lagged)	0.872**	-1.936	-0.736	-3.615*	-1.378	7.178
	(2.11)	(1.60)	(0.63)	(2.17)	(1.03)	(1.12)
ENTR_INNO (2y lagged)	0.003	-0.011	0.018	-0.114**	0	0.007
	(0.56)	(1.29)	(0.53)	(2.30)	(0.06)	(1.31)
INNO/NPE * ENTR_INNO (both 2y lagged)		7.873**		13.545***		-16.253
		(2.66)		(3.40)		(1.29)
ENTR_EMP (1y lagged)	-0.044**	-0.040**	-0.046	-0.053	-0.039**	-0.037**
	(2.67)	(2.38)	(1.06)	(0.90)	(2.82)	(2.56)
year	0.800**	0.699*	0.925	0.692	0.572**	0.590**
	(2.17)	(2.01)	(1.32)	(1.07)	(2.58)	(2.75)
year²	0.000**	0.000*	0.000	0.000	0.000**	0.000**
	(2.17)	(2.01)	(1.32)	(1.06)	(2.58)	(2.75)
_cons	-803.722**	-701.719*	-930.145	-695.108	-574.378**	-593.032**
	(2.18)	(2.01)	(1.33)	(1.07)	(2.59)	(2.76)
# observations	177	177	92	92	85	85
# groups	22	22	11	11	11	11
F statistic	38.62	51.13	39.55	44.54	29.68	149.8
R-squared Within	0.49	0.53	0.48	0.54	0.77	0.79
R-Squared Between	0.54	0.58	0.19	0.24	0.3	0.23

* $p<0.1$; ** $p<0.05$; *** $p<0.01$

The dynamics behind these observed effects could be explained among others by referring to a mix of economies and diseconomies of scale (Brock & Evans, 1989). The observation of an entrepreneurial innovation premium could be attributed to the higher likelihood that patents introduced by small businesses will be high impact ones, making the average small firm patent more technically and thus more economically important. This finding complies with Baumol's (2004) assignment of different roles to small and large firms in innovation systems with the former being relatively better at the introduction of radical new technologies and the latter in perfecting those by incremental improvements. The observed discount observed on the entrepreneurialisation of employment suggests that in the non-innovation-related aspects of business operations the economies of scale outweigh the diseconomies of scale. This observation counters earlier findings underlining the increasing importance of non-technologically oriented scale diseconomies that result from growing markets valuing specialized products, increasing advantages to flexibility in a globalized world, the rising availability of educated labour to recruit from and decreasing standard fixed costs of running a business (Brock & Evans, 1989).

Separate results for countries tagged by the EC as innovation leaders and laggards further reveal some of the potential deeper dynamics behind this. Not surprisingly, the innovation leaders turn out to be the driving force behind the productivity premiums to technological innovation in general and entrepreneurial innovation. The former and latter can be seen as highly intertwined: established knowledge-based economies possess the critical mass that is necessary to produce knowledge that matters. Knowledge stock growth in turn increases the potential for spill-overs of various ideas to entrepreneurs. On top of that, local rivalry between high-tech entrepreneurial ventures capturing the same localized knowledge flows increases their respective efficiency (Furman et al., 2002). The laggard countries appear to be the

driving force behind the productivity discounts associated with small firm employment share growth. The distinct geographic origins of the premium effect on entrepreneurial innovation and discount effect on entrepreneurial employment confirm the heterogeneous nature of the European economic landscape. Relying on Porter et al.'s (2002) framework of economic development to explain differences between split dataset results one could claim that it consists of less developed countries in a 'preliminary' efficiency-driven stage and more advanced countries in the 'final' innovation-driven stage (Porter et al., 2002; Acs et al., 2008).

In a complementary attempt to explain the additional explanatory power of entrepreneurial innovation in general we refer to the increasing disjunction between patents as measure of innovation and productivity in large firms: the availability of in-house IP departments increase their propensity to patent low-value inventions and tax optimization strategies applied by multinationals blur the value of license fees as proxy for added value.

Future research is necessary to further disentangle the mechanics behind the observed effects. Measurement of knowledge spill-overs could help to provide insights about their nature, origins and the direction in which they are heading. Adding proxies capturing the distinct drivers of scale diseconomies is another potential direction for future research. Further inquiry is also needed to list the policy implications of our findings.

References

Acs, Z. J., Desai, S., & Hessels, J. (2008). Entrepreneurship, economic development and institutions. *Small Business Economics,* 31, 219-234.

Audretsch, D. B. (2001). The dynamic role of small firms: evidence from the US, *Small Business Economics,* 18 (1/3), 13-40.

Baumol, W. J. (2004). Entrepreneurial enterprises, large established firms and other components of the free-market growth machine, *Small Business Economics,* 23, 9-21.

Brock, W. A. & Evans, D. S. (1989). Small business economics, *Small Business Economics* 1 (1), 7-20.

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

European Commission (2005). The new SME definition: user guide and model declaration, *Enterprise and industry publications,* European Commission.

European Commission (2015). Innovation Union Scoreboard 2015.

Eurostat (2014). Patent statistics at Eurostat: mapping the contribution of SMEs in EU patenting.

Fraunhofer Institute, Systems and Innovation Research (2003). Patents in the service industries. Final report European Commission Contract No. ERBHPV2-CT-1999-06.

Furman, J. L., Porter, M.E., & Stern, S. (2002). The determinants of national innovative capacity, *Research Policy,* 31, 899-933.

Nadiri, I. (1993). Innovations and technological spillovers, NBERWorking Paper 423.

Porter, M., Sachs, J., & McArthur, J. (2002). Executive summary: Competitiveness and stages of economic development. In M. Porter, J. Sachs, P. K. Cornelius, J. W. McArthur, & K. Schwab (Eds.), *The global competitiveness report 2001-2002* (pp. 16-25). New York: Oxford Univ.Press.

Romer, P. M. (1986). Increasing returns and long run growth*, Journal of Political Economy,* 94, 1002-1037.

Schumpeter, J. A. (1942). *Capitalism, socialism and democracy,* Harper and Row: New York.

Solow, R. M. (1956). A contribution to the theory of economic growth, *Quarterly Journal of Economics,* 70, 65-94.

Ulku, H. (2004). R&D, innovation, and economic growth: an empirical analysis, IMF Working Paper 04/185.

Veugelers, R. (2009). A lifeline for Europe's young radical innovators, Bruegel Policybrief 1.

Wong, P.K., Ho, Y.P., and Autio, E. (2005). Entrepreneurship, innovation and economic growth: evidence from GEM data, *Small Business Economics,* 24, 335-350.

Chemistry research in India: A bright future ahead

Swapan Deoghuria[1*] Gayatri Paul[2] and Satyabrata Roy[3]

[1]ccsd@iacs.res.in * *Corresponding Author*
Scientist-III, Indian Association for the Cultivation of Science, Jadavpur, Kolkata – 700032 (India)

[2]*libgp@iacs.res.in*
Sr. Doc. Assistant, Indian Association for the Cultivation of Science, Jadavpur, Kolkata – 700032 (India)

[3] *saturoy@gmail.com*
Guest Faculty, LIS Department, Jadavpur University, Jadavpur, Kolkata – 700032 (India)

Introduction

Chemistry is the most preferred research area among Indian scientists for quite some time in terms of total number of publication, global share, visibility and citation impact are concerned. Growth rate of India in chemistry research area is more than that of global growth rate as evidenced from the data covered in Web of Science database (WoS). The trend of research output in chemistry clearly indicates that India is steadily putting stiff challenge to traditionally established countries like Japan and Germany and even surpassed them in 2014 to acquire 3rd position in global ranking. From this study we predict that India will grow further in chemistry research area and even can put challenge to USA and China in long run.

The output and trend of science & technology (S&T) research in India are of considerable interest to scientometricians from all over the world for quite some time. Gupta and Dhawan (2009), Glänzel and Gupta (2008) and Gunasekaran, Batcha and Sivaraman (2006) have studied different aspects of S&T research in India.

Methodology

Data sources and processing

All bibliometric data have been extracted from WoS Core Collection of Thomson Reuters till April 30, 2015. The period for publication activity has been taken for six years (2009-2014) as findings till 2008 are available in literature.

Results and Discussions

In chemistry research area a total 1,045,343 number of papers has been published during the period 2009-2014. USA and China are leaders in this field in terms of number of publications with global share of 22.502% and 20.792% respectively. India is at 5th position with global share of 5.767%. Chemistry research output of ten most productive countries excluding USA and China in terms of global share has been shown in Figure 1. India's growth is very steady during this period and acquired 3rd position in 2014 followed by USA and China, with global share of 6.456%. India has published maximum number of research papers in Chemistry compared to other research areas and its global share in chemistry research has been increased steadily during 2009 to 2014.

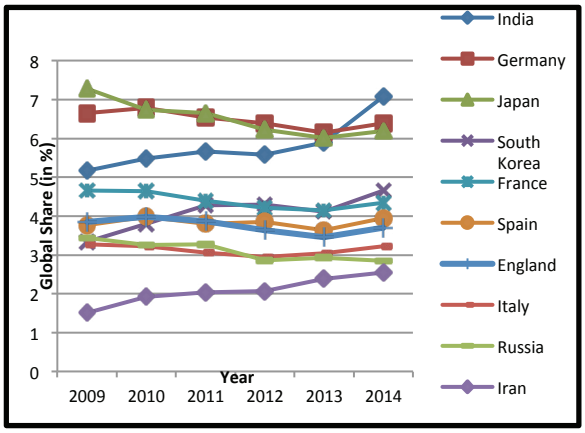


Figure 1. Global share of countries in chemistry.

It is evident from Figure 1 that global share of Japan has been decreased during 2009-2014 and its positions in global ranking have been fallen from 3rd position in 2009 to 5th position in 2014. Global share of Germany in Chemistry research has been decreased slightly during this period but Germany has managed to keep its position at 4th during the entire period. South Korea and Iran have increased their research output in chemistry steadily in terms of global share during this period. Research output of other countries (France, England, Spain, Italy and Russia) shown in this Figure are comparable to each other in chemistry and they are placed in between 7th to 11th positions during this period.

Table 1 shows India's ranking in major research areas covered in WoS during 2009-2014. In terms of number of publications and global share, India's performance is the best in Chemistry.

In Table 2 we have shown the h-index and average citation per article in chemistry during 2009-2012. We see that h-index and average citation per article are comparable with that of Japan and Germany.

Table 1. India's Position in major research areas in terms of global share.

Research Areas	2009	2010	2011	2012	2013	2014
Physics	10	9	8	8	7	**7**
Chemistry	5	5	5	5	5	**3**
Materials Science	7	6	6	6	**5**	6
Engineering	11	12	11	6	**4**	6
Computer Science	12	12	9	**3**	4	11
Biochemistry Molecular Biology	12	11	11	11	10	**9**
Neuroscience Neurology	18	17	17	16	**16**	17

Table 2. Comparison of citation and h-index of chemistry publications during 2009-2012.

Countries	2009		2010		2011		2012	
	h- index	Avg Citation	h- index	Avg Citation	h- index	Avg Citation	h- index	Avg Citation
India	83	11.54	78	10.39	66	8.41	57	6.49
Japan	106	15.83	97	13.81	89	11.63	66	7.94
Germany	118	19.86	125	19.29	95	14.04	74	9.94

Conclusions

This study clearly indicates the trends in chemistry research during 2009-2014 for most productive countries in terms of number of publications and global share. It is evident from the results that India has done remarkable progress in chemistry research area during this period. One of the reasons for this progress is that quite a few key persons in science policy makers in India are having chemistry background. Indian scientists working in the field of chemistry are more focused and recognized worldwide as many of them have been awarded TWAS prize and fellowship, FRS, and other distinguished international fellowships and medals. Strong collaboration between India and other countries in chemistry research is worth mentioning as 10,941 numbers of papers out of total 60,285 are published in collaboration. As a traditional subject, most of the Indian universities teach chemistry and around 40% of total publications is contributed by the universities. Research laboratories also get a steady flow of trained students with chemistry background from universities. Looking at the distribution of the publications to the institutes we see that CSIR laboratories publish most (11,037) followed by IITs (7,382) in chemistry. Some of the most productive laboratories in chemistry research in India are BARC (2,394), IICT (2,210), IISc (2,065), IACS (1580) and NCL (1,508). Prominent universities in chemistry research are JU (1,262), DU (1,182) and BHU (1,136). We see that there is almost no role of industries as per the funding of research is concerned in the field of chemistry in India. CSIR, DST and UGC are the major sponsors in chemistry research in India. As per the topic or subject category is concerned where Indian scientists publish more, we see Physical chemistry is the most focused (29%) followed by Organic (20%), Inorganic (11%), Analytical (10%), Applied (7%), Nanoscience (6%) and Atomic-Molecular (5%) respectively. The bright side of chemistry research in India is also reflected in the number of patents granted in this subject area. From Derwent Innovations Index of WoS, we see that out of total 462 numbers of patents granted to Indian innovators during 2009-2014, 330 numbers i.e. 71% are in the field of chemistry. Interestingly, DRDO, India holds most (79%) of the patents. The picture is not much different in Indian patent database (http://ipindiaservices.gov.in/publicsearch/), where we see 4,801 numbers of patents (i.e. 37%) have been granted in chemistry research area out of total 12,982 patents granted in all fields during 2009-14. India has a large consumer base. As a result chemical industries in different sectors like fertilizer, pesticide, plastic, paint, petro-chemical, medicine, cosmetics and health care products are thriving in India. So career as research scientist in chemistry is attractive for better placement in the R&D labs of those industries. India's contribution in chemistry research has been recognized by ACS and designated IACS, Kolkata on 15/12/1998 as International Historic Chemical Landmark for C V Raman and the Raman Effect.

References

Glänzel, W & Gupta, B. M. (2008). Science in India. A bibliometric study of national and institutional research performance in 1991-2006. *Proc. WIS.*

Gunasekaran, S., Batcha, M. S. & Sivaraman, P. (2006). Mapping chemical science research in India: A bibliometric study. *Annals of Library and Information Studies, 53*, 83-95.

Gupta, B. M. & Dhawan, S. M. (2009). Status of India in science and technology as reflected in its publication output in Scopus Int. databases, 1996 – 2006. *Scientometrics, 80*(2), 473-490.

Main Institutional Sectors in the Publication Landscape of Spain: The Role of Non-profit Entities

Borja González–Albo[1], Javier Aparicio[1], Luz Moreno-Solano[2], María Bordons[2]
[1] *borja.gonzalezalbo@cchs.csic.es; javier.aparicio@cchs.csic.es*
Transversal Support Research Unit (UTAI), Centre for Humanities and Social Sciences (CCHS), Spanish National Research Council (CSIC). Albasanz 26–28, 28037 Madrid (Spain)

[2] *luz.moreno@cchs.csic.es; maria.bordons@cchs.csic.es*
ACUTE Group, IFS, Centre for Humanities and Social Sciences (CCHS), Spanish National Research Council (CSIC). Albasanz 26–28, 28037 Madrid (Spain)

Introduction

The study of national efforts in R&D by institutional sector is a matter of great concern because sectors differ in their main activities, accounting systems, orientation towards research and type of R&D (OECD, 2003). However, bibliometric analyses at the level of institutional sectors are not very common because the assignation of centres to sectors is not free of difficulties and the resulting sectors may entail a certain degree of heterogeneity. The role of institutional sectors in the scientific activity of countries, either for the total country (Godin & Gingras, 2000; Moya et al., 2013) or in a given field (Lander, 2013), has been analysed in the literature, although studies dealing with specific sectors such as universities or companies are much more frequent.

In most countries, main institutional sectors in publications include universities, hospitals and public research centres, while papers from non-profit entities (NPE) are usually scarce. Although this applies in Spain, an impressive increase in papers from NPE has been observed in the last fifteen years. This paper aims to analyse the research performance of non-profit entities in Spain with regard to activity, impact and collaboration; to locate them in the national context; and to identify main types of active organisations.

Methods

Spanish publications (original articles and reviews), hereafter papers, covered by Web of Science (WoS, 2000-2011), search strategy CU=Spain and PY=2000-2011, are analysed. Six institutional sectors are identified in all addresses through a semi-automatic process (Morillo et al., 2013) followed by a manual revision to assess validity: companies, health sector, non-profit entities, public administration, public research centres and university. A full counting method is used.

The impact of publications is analysed through the percentage of papers in first quartile journals within each field (%Q1), normalised position (NP) (Bordons & Barrigón, 1992), relative impact factor (RIF), % non-cited papers and citations relative to country average (RC) (three-year citation window). The orientation of sectors towards collaborative research is explored through the number of authors per paper, number of institutions per paper and collaborative pattern (percentage of papers with a single institution, percentage of papers with national collaboration, percentage of papers with international collaboration). An in-depth analysis of NPE is carried out. The NPE's activity index (AI) in ten broad thematic areas is obtained to gain insight into the specialisation profile of these entities as compared to Spain.

Results

Main institutional sectors in Spanish papers in WoS (2000–2011) include university (66%), public research organisations (22%) and the health sector (18%). Non–profit entities amount to 10% of the papers, and show the highest increase during the period (3% of the country output in 2000 vs. 18% in 2011). This sector shows high specialization in Biomedicine (AI=1.59) and Clinical Medicine (AI=1.67). Collaboration in NPE is above the country average in terms of team size (11 vs. 8), number of institutions per paper (5 vs. 3) and share of collaborative papers (91% vs. 68%). NPE show also the highest shares of both nationally and internationally co-authored papers (75% vs. 41% and 45% vs. 40%, respectively). NPE display the highest percentage of papers in high-quality journals and the highest impact through relative citations (Table 1).

From the inspection and categorization of the NPE, the following organisational types emerge: foundations (50.3%), research networks (24.6%), consortia (16.0%), research management entities (12.2%), associations (6.5%), and scientific parks (1.0%). The highest increase during the period corresponds to research management entities and research

networks. Research management entities stand out because of their high figures in both the percentage papers in high impact factor journals and relative citations (Table 2).

Research management entities show the lowest proportion of papers with a single institution (2%), a high share of papers with national (89%) and international collaboration (68%), and the highest average team size. The highest share of papers in Q1 journals is observed for co-authored activity between national and foreign partners for all sectors except associations and research networks.

The specialization of NPE varies according to the organisational type: Biomedicine and Clinical Medicine for networks, consortia and foundations; Physics for research management entities; Biomedicine and Chemistry for scientific parks; and Engineering for associations.

Conclusions

The in-depth analysis of the NPE in Spain shows the rising trend of different organisational types which differ according to the field and respond to specific strategic procedures to manage research (creation of foundations in the context of medicine, networks for clinical research, scientific parks to link basic and applied research in the university context, etc.). Interestingly, some of these organisational types (research networks, consortia, parks) include cross-sector and cross-discipline collaboration which is supposed to lead to major discoveries in science and even to radical innovation. Collaboration in the context of the structured and stable framework provided by these organisational forms is more effectively enhanced than through occasional collaborative projects. Our data indicate the success of these emerging organisations in supporting/conducting high impact research.

Acknowledgements

The financial support of the Spanish Ministry of Science and Innovation (CSO2011-25102) is acknowledged.

References

Bordons, M. & Barrigón, S. (1992). Bibliometric analysis of publications of Spanish pharmacologists in the SCI (1984-89). Part II. Contribution to subfields other than "Pharmacology and Pharmacy", *Scientometrics*, 25 (3): 425–446.

Godin, B., & Gingras, Y. (2000). The place of universities in the system of knowledge production. *Research Policy*, 29 (2), 273–278.

Lander, B. (2013). Sectoral collaboration in biomedical research and development. *Scientometrics*, 94 (1), 343–357.

Morillo, F., Aparicio, J. González–Albo, & B., Moreno, L. (2013). Towards the automation of address identification. *Scientometrics*, 94 (1), 207–224.

Moya–Anegón, F., Chinchilla–Rodríguez, Z., Corera–Álvarez, E., González–Molina, A., López–Illescas, C., & Vargas–Quesada, B. (2013). *Indicadores bibliométricos de la actividad científica española 2010*. Madrid: FECYT.

OECD (2002). *Frascati Manual*. Paris: OECD.

Table 1. Number of papers and impact indicators by institutional sector in Spain (WoS 2000-2011)

	No. Papers	NP	%Q1	%Non cited papers	RC	RIF
Universities	271399	0.66	47.93	23.45	0.85	0.89
Public Research Centres	91095	0.74	62.41	12.94	1.31	1.24
Health sector	74337	0.59	39.66	21.32	1.20	1.16
NPE	41605	0.74	62.59	10.56	1.75	1.57
Public Administration	17238	0.66	49.04	20.65	1.01	0.96
Companies	15682	0.63	43.72	22.15	0.81	0.84

Table 2. Number of papers and impact indicators of the NPE by organisational type (WoS 2000-2011)

	No. Papers	NP	%Q1	%Non cited papers	RC	RIF
Foundations	20934	0.76	65.50	9.71	1.82	1.67
Research Networks	10249	0.75	63.16	7.18	1.83	1.74
Consortia	6651	0.73	60.83	9.88	1.69	1.55
Research Management Entities	5074	0.81	76.47	6.42	2.71	1.96
Associations	2692	0.66	47.73	20.84	0.90	0.94
Scientific Parks	310	0.76	66.11	8.71	1.21	1.55
Other NPE	1204	0.60	35.35	27.99	0.75	0.76

Reform of Russian Science as a Reason for Scientometrics Research Growth

Andrey Guskov

guskov@spsl.nsc.ru
The State Public Scientific Technological Library of Siberian Branch of the Russian Academy of Science, Voshod Str. 15, Novosibirsk (Russia)
Novosibirsk State University, Pirogova Str. 2, Novosibirsk (Russia)

Introduction

After the USSR had fallen down in 1990, there was a steady stagnation of Russian science for fifteen years. Iron curtain that separated soviet researchers from the international science disappeared, but research funds sharply decreased due to the economic problems. As a result, the number of publications registered in Web of Science, stayed between 30 000 and 34 000 per year. Thus, Russian science moved from the group of leading countries to the second dozen.

Restoration of Russian Science started in 2006 after government had introduced a new model of the research process. Essential part of the model was wide application of the formal scientific results assessment. This approach triggered a rapid growth of scientometrics publications written by mathematicians, physicists, philosophers and others. The main goal of this paper is to make a review of new Russian scientometrics landscape, which could help to determine its strengths and weaknesses and launch new collaborations.

Method

In this paper basic set of scientometric articles produced by Russian scientists is analysed. It consists of two periods: 1988-1999 and 2000-2014. The data for the first part (99 publications) was extracted from Russian Institute for Scientific and Technical Information database, abstract journal "Informatics" (Penkova, O. & Tyutyunnik V., 2011) Publications from 2000 until 2014 were requested from Russian Science Citation Index (national bibliometric database) by using context search with terms "bibliometric", "scientometric", and "webometric" (in Russian) in titles and annotations.

For every article in this set we identified topic category according to its title, annotation and, in some cases, full text. Afterwards, we analysed the distribution and dynamics of the categories and of the whole set.

Dynamics of Russian scientometric researches

Noteworthy, scientometrics in Russia has very meaningful historical background. It was Russian philosopher and mathematician V. Nalimov, who in 1969 introduced the term "Scientometrics" in his famous book. In 1973 Marshakova and Small simultaneously introduced co-citation analysis, which is used for research front findings now. Dutt, Garg & Bali in 2003 analysed fifty volumes of journal Scientometrics during 1978 to 2001 and examined the distribution of the output of different countries. According to their paper, former USSR contributed 59 of 1317 articles that are emphasized on history of science, theoretical studies and scientometrics distribution. Despite these go-ahead results, scientometric researches became a trend in Russia only after 2006 (Fig.1).

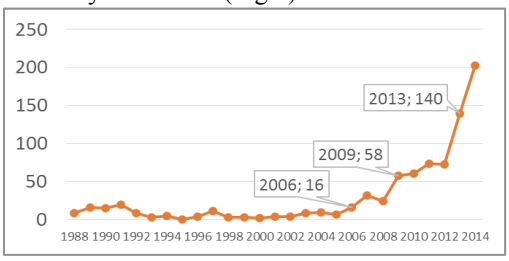


Figure 1. Dynamics of scientometric publications in Russian journals.

There are three sharp increases at Fig 1: in 2006, 2009, and 2013. The first growth in 2006 relates to the reformation of salary system, which implied significant dependency of the payment bonuses upon publication scores for every single scientist. Facing this new challenge, a number of researchers considered its fairness; some of them noticed the helpfulness of the bibliometric methods and started to apply it for their subject area. The second wave started in 2009th after the end of the salary system reformation. From that moment, every researcher became financially interested in improving his scientometric indicators. Research society had to analyze these changes, thus we can observe sharp increase in 2009th at Fig 1.

Despite the rapid growth before, in 2013th the number of scientometric publications had doubled. The reason is clear: in May 2012, President of Russia V.V. Putin proclaimed that the fraction of publications of Russian researches indexed by Web of Science in 2015th has to be greater than 2.44%. This was quite a big challenge for national science, because it literally meant that the annual number of articles has to be increased from 32-33 thousands in 2010-2011 to 46-50 in the next 3 years. The

reasons, the ways and the possibilities of that breakthrough were the main topics for discussion over the year. After that, in June 2013 another dramatic event occurred: restructuring of the Russian Academy of Science (RAS), headquarters of fundamental sciences. This tough stage was accompanied by criticism of the Academy for low scientometric indicators. Unfortunately, scientometrics has been used as an instrument for a radical transformation of management of Russian science.

Directions of researches

We defined 16 categories and analyzed the articles distribution (Fig.2). 33% of researches were devoted to a specific subject area investigation. It is followed by: development and applying of indicators (13%), general discussions about scientometrics and its place in research management (11%), impact-factors and journal improvement issues (7%), positions of Russian science in a global scope (6%). According to our estimates, from 50% to 75% of publications were made using bibliometric methods, principally in categories: "Subject areas", "Journals", "National science", "Dissertations", "Regional research", "Leading scientists research", "Science in HEI", "Conferences", "Organizations", "Collaborations", "Patents".

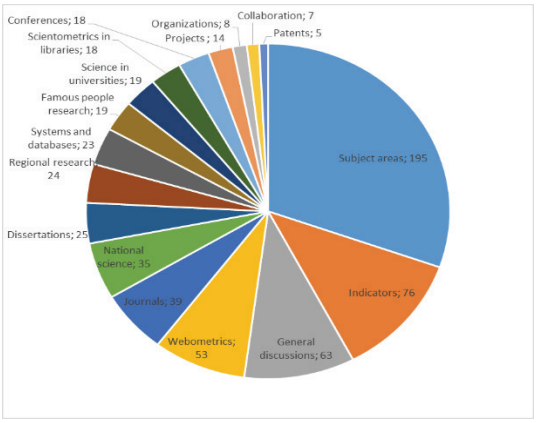


Figure 2. Distribution of scientometric researches by categories (number of publs.)

We determined the most developing categories and analyzed the dynamics. The main contribution to publication rise, shown at Fig.1, was made by "Subject area" category from 2007 to 2012. The second contributing category "Indicators" contains a number of articles about publications and citations amount, impact factor and Hirsh index. The third category supports general scientific discussion about scientometrics, started in 2009. Three more categories significantly increased in 2013: "Journals", "Science in universities", "Systems and databases".

Conclusion

Figure 1 can be thought of as an indirect measure of the influence of the State on Russian Science. Indeed, there was a lack of scientometricians and poor scientometric publication activity in Russia before 2006th, the very beginning of reformation. The following alterations made many researches slow down or suspend what they had been doing before and start making their own scientometric investigations. The more severe were the changes, the more scientists were influenced. Furthermore, it seems there were no other reasons for the mentioned breakthrough. At first glance, scientometrics is supposed to benefit from it. That would be so, excepting two facts. First, concerning scientometrics as an instrument of reformation, many scientists consider it primarily as a stick for punishment and do not trust it. This creates quite a negative environment for further development, but this story has already happened. "When a system of assessing and funding researchers was introduced in South Africa, there were cases when scientists attacked scientometrics…" (Pouris, 1994). Second, the most of the scientometric researches, which were published in Russia the last years, relate to one of the groups: 1) Position of the scientometrics and its indicators in the processes of the management of Russian science. 2) Bibliometric researches of science disciplines and Russian science as a whole. 3) Bibliometric and webometric researches of various sources of publications: journals, organizations (incl. universities), famous scientists, conferences, projects, dissertations sets and so on. Since those three groups include up to 90% of publications, there is not much space left for more complicated and go-ahead researches, such as collaboration studies, research fronts detecting, R&D cycle analysis, altmetrics, society impacts, etc. At the moment, scientometrics in Russia remains the "product for internal use" mostly. Still, we expect the internalization of this research field and the increase of the visibility of Russian publications worldwide.

Acknowledgements

The author acknowledges Denis Kosyakov, Irina Selivanova and Mikhail Tsentalovich.

References

Dutt, B., Garg K. & Bali, A. (2003) Scientometrics of international journal Scientometrics. *Scientometrics*, Volume 56, No. 1, pp 81-93. DOI: 10.1023/A:1021950607895

Penkova, O. & Tyutyunnik V. (2001) Informetrics, scientometrics and bibliometrics: the sciento-metrics analysis of current state. *Vestnik Tomskogo gosudarstvennogo universiteta.* 6(1) 86-87.

Pouris, A. (1994) Is scientometrics in a crisis? *Scientometrics*, Volume 30, Nos 2-3, pp 397-399. DOI: 10.1007/BF02018111

Leadership among the Leaders of the Brazilian Research Groups in Marine Biotechnology

Sibele Fausto[1] and Jesús P. Mena-Chalco[2]

[1] *sifausto@usp.br*
University of São Paulo, Rua da Biblioteca, s/n, Complexo Brasiliana,
São Paulo, SP, CEP 05508-050 (Brazil)

[2] *jesus.mena@ufabc.edu.br*
Federal University of ABC, Av. dos Estados, 5001, Santo André, SP, CEP 09210-580 (Brazil)

Introduction

The Marine Biotechnology (MB) research area is gaining increasing relevance in Brazil. Its analysis is a challenge owing to the inherently multidisciplinary nature, and the study of research groups (RGs) may support this work. The task of analysing RGs is facilitated in Brazil, which has a national source gathering the country's RGs, maintained by the National Council for Scientific and Technological Development (*Conselho Nacional de Desenvolvimento Científico e Tecnológico* - CNPq): the Directory of Research Groups of the Lattes Platform (*Diretório dos Grupos de Pesquisa da Plataforma Lattes,* http://lattes.cnpq.br/web/dgp), with information from RGs related to: i. institutional headquarters; ii. Research Group name; iii. First leader name, iv. Second leader name (if any), and v. Predominant area. This source allows automatic data extraction already made available by research groups, allowing for full and systematic exploitation. This work aims to present first findings from exploitation on research groups in MB existing in Brazil registered in the Directory of RGs of the Lattes Platform, checking the collaboration networks formed by the leaders of these groups, mainly highlighting the natural influence that leaders have on other peers, meaning a leadership, focusing on research groups through the topological properties of networks with the use of Social Network Analysis (Abbasia, Wigand & Hossain, 2014), in order to behold their evolution and the role of the RGs' leaders in MB in Brazil and testing if it is possible to establish a relationship between the degree of leadership of the leaders considering topological information from networks.

Methods

This initial approach is focused on three points: 1. networks characterization in number of RGs involved, the active institutions and their location, and the dominant areas in multidisciplinary research; 2. description of the dynamic aspect of the network formed by these RGs through its evolution over the last 15 years, distributed in three five-year periods; and 3. determination of the "degree of leadership" of these networks' leaders, as measured by AuthorRank indicator, which is a numerical value that indicates the impact of a member in collaboration graph. This measurement is similar to PageRank for directed graphs (with weights) (Liu et al., 2005). Thus, the aim was to consider this indicator as an attribute of the leadership for the leaders of these RGs in the analyzed period.

Data collection and analysis

First, the MB research groups were identified by search using 37 MB terms raised in the related literature. Following, it was obtained data related to RGs such as institutions involved, 1st Leader name, and Main Area, allowing identify the Lattes ID (researcher identification number registered in the Lattes Platform) of the groups' leader. Second, we used scriptLattes tool (Mena-Chalco & Cesar Junior, 2009) in order to extract information associated with all the investigated leaders during the period of 15 years (1999-2013). We obtained data from the scientific production of each leader related to total articles, books, book chapters, and conference papers. For data analysis, we consider the professional addresses recorded for each leader to obtain the geographic location of each group through Google Maps tool. We obtained lists of full papers (solely) of the groups' leaders published in journals, and with scriptLattes tool we identify all publications in co-authorship. In addition, there were obtained the endogenous networks (internal collaboration) of the leaders. The AuthorRank was calculated for each actor. This indicator is commonly used for measuring the impact of members of an academic collaboration network (Liu et al., 2005). Our analysis was outlined considering four time periods: A global period (1999-2013) and three five-year periods: 1999-2003, 2004-2008, and 2009-2013. This division into different periods allows to study distinct topological characteristics of the network and its evolution.

Results

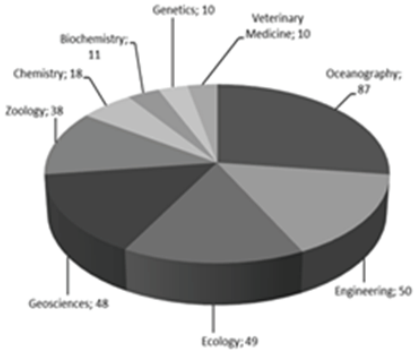


Figure 1. Main subject areas of the Brazilian research groups in Marine Biotechnology

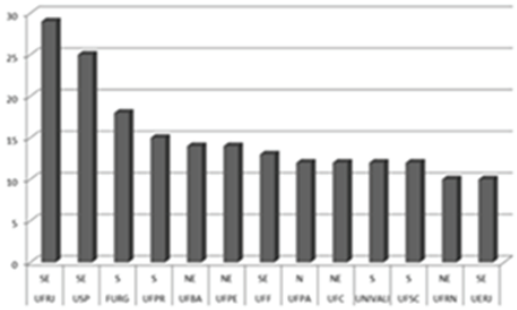


Figure 2. Brazilian institutions with over ten research groups in Marine Biotechnology

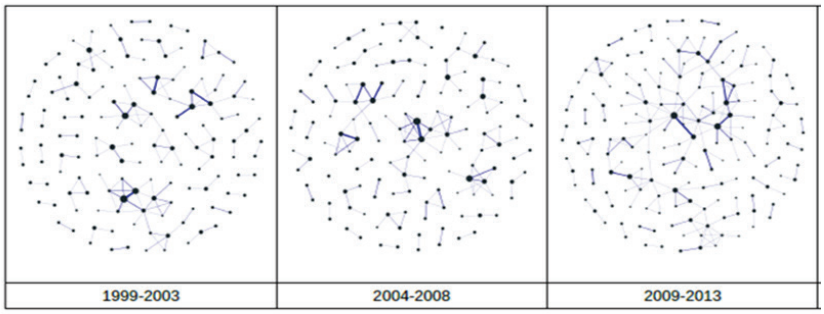


Figure 3. Co-authorship networks among leaders associated with the Brazilian research groups in Marine Biotechnology

Table 1. AuthorRank of the Leaders of the Brazilian research groups in Marine Biotechnology

Author Rank	Leader	Institution/Region
4.10	Teixeira, VL	UFF/SE
3.59	Colepicolo Neto, P	USP/SE
3.46	Rörig, LR	UFSCar/SE
3.33	Pereira, RC	UFF/SE
2.94	Pinto Jr, E	USP/SE
2.75	Mantelatto, FLM	USP/SE
2.67	Amado Filho, GM	JBRJ/SE
2.55	Sampaio, LAN	UFRN/NE
2.34	Bianchini, A	UFRN/NE
2.33	Berlinck, RGS	USP/SE

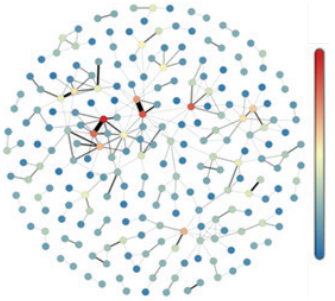


Figure 4. AuthorRank of the Leaders of the Brazilian research groups in Marine Biotechnology: co-authorship network

Discussion and conclusion

There are 402 RGs working in one or more topics related to the MB field from 34 different subject areas, main ones showed in Figure 1. RGs are from 110 institutions geographically concentrated along the Brazilian coast (South and southeast prevailing in number of institutions and research groups – Figure 2). We identified the leadership of the ten most active researchers in the co-authorship networks, with AuthorRank varying between 2.33 and 4.1 (Table 1). It was observed that there is a systematic increase in academic interactions during the considered period (Figure 3) and that academic leadership is not uniform among the leaders (Figure 4). The task of characterizing the emerging area of research in MB has grown in importance in Brazil, and this work relates to this issue.

References

Abbasia, A., Wigand, R. T. & Hossain, L. (2014). Measuring social capital through network analysis and its influence on individual performance. *Library & Information Science Research*, 36, 66–7.

Liu, X., Bollen, J., Nelson, M. L., & van de Sompel, H. (2005). Co-authorship networks in the digital library research community. *Information processing & management*, 41(6), 1462-1480.

Mena-Chalco, J. P. & Cesar Junior, R. M. (2009). ScriptLattes: an open-source knowledge extraction system from the Lattes platform. *Journal of the Brazilian Computer Society*, 15(4), 31-39.

An Empirical Study on Utilizing Pre-grant Publications in Patent Classification Analysis

Chung-Huei Kuan[1] and Chan-Yi Lin[2]

[1] *maxkuan@mail.ntust.edu.tw*
National Taiwan University of Science and Technology, Graduate Institute of Patent Research, No. 43 Sec. 4 Keelung Rd., Taipei City, 106 (Taiwan, R.O.C.)

[2] *u9703220@gmail.com*
National Taiwan University of Science and Technology, Graduate Institute of Patent Research, No. 43 Sec. 4 Keelung Rd., Taipei City, 106 (Taiwan, R.O.C.)

Abstract
Patent classification analyses are usually conducted using issued patents. Issued patents however suffer lengthy examination and the derived analytic results reflect R&D activities occurring considerable time in the past. The only option for an analyst to reduce such observational time delay is to use the so-called pre-grant publications (PGPubs) that are open to public 18 months after patent applications are filed. The PGPubs and their corresponding issued patents are both assigned classification symbols. If the two sets of symbols are very different, using patent classification analysis on PGPubs to observe R&D activities is dubious. This study therefore compares the United States Patent Classification (USPC) symbols assigned to about 235,000 pairs of U.S. utility patents issued in 2012 and their PGPubs in three ways, each corresponding to an approach of a conventional patent classification analysis: (1) considering only the class codes of the main classification symbols; (2) considering only the main classification symbols; and (3) considering both main and auxiliary classification symbols. The study finds that only the class codes of the PGPub main classification symbols are reliable enough for patent classification analysis as there are about 78% of the PGPubs have identical class codes as their corresponding issued patents.

Conference Topic
Patent analysis

Introduction
A patent application is classified during its prosecution process based on its inventive content by an examiner and one or more classification symbols are assigned in accordance with a standard scheme such as International Patent Classification (IPC), Cooperative Patent Classification (CPC), U.S. Patent Classification (USPC), etc. Patent classification analysis (PCA) is a popular practice by patent analysts using the patent classification symbols, and it is so popular that, to the authors' knowledge, all commercial patent analytic systems/services, such as Thomson Innovation® and WIPS Global®, have various types of PCA built-in.

A common type of PCA is to investigate the R&D focuses of an entity (i.e., a company, an institute, a country, a technical field, etc.). An analyst gathers the patents affiliated with the entity, collects the classification symbols assigned to these patents, counts the number of times each classification symbol is assigned to these patents, and usually produces a diagram such as a histogram, a heat map, etc., to visually manifest the assignment frequencies of the classification symbols. By observing the diagram, the analyst then claims that the entity has its R&D focused in a few technical areas denoted by the most frequently assigned classification symbols.

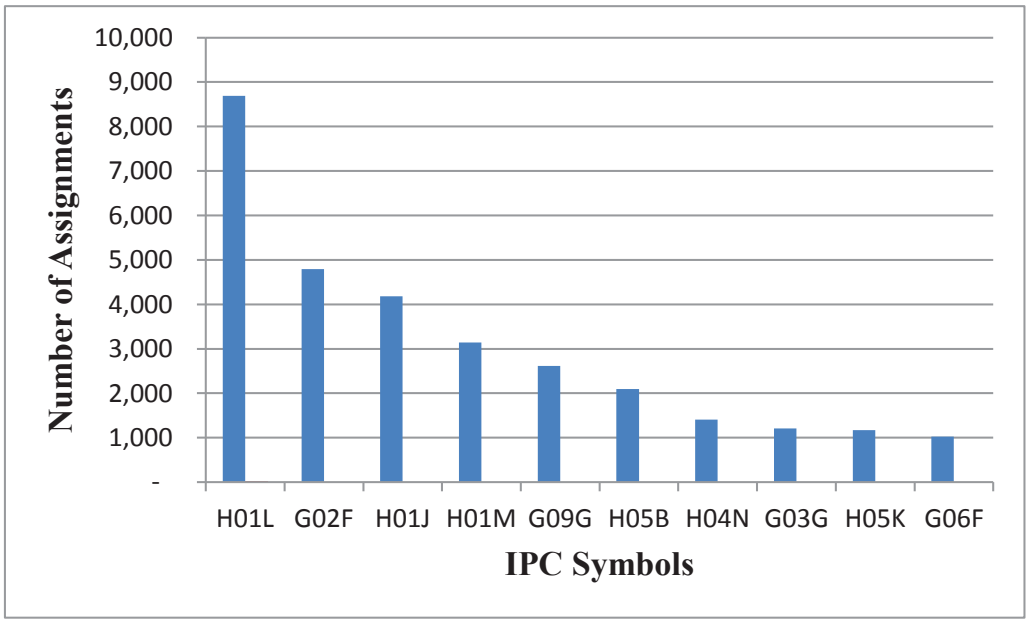


Figure 1. A sample histogram from a fictitious PCA.

A sample histogram from a fictitious PCA using IPC symbols for a company is shown in Figure 1. As illustrated, the company is considered to have its R&D effort mainly focused in the field Semiconductor Devices denoted by the most frequently assigned IPC symbol H01L.

Other than the real-life application described above, patent classification symbols are considered as a viable source of technological information by researchers, and various types of PCA have been proposed in the literature. To mention just a few, the number of different classification symbols assigned to an entity's patents is used as a proxy to the entity's technological diversity (cf. Lerner, 1994), the co-classification of patents (i.e., patents assigned one or more identical classification symbols) is used to investigate the linkage among technologies (cf. OECD, 1994), or the relationships among organizations (cf. Leydesdorff, 2008). There are also studies investigating the technological relatedness of two entities using the classification symbols assigned to their patents (cf. Jaffe, 1986; 1989). In addition, the classification symbols of a patent's forward and backward citations are used to evaluate the patent's "generality" and "originality" (cf. Henderson, Jaffe, & Trajtenberg, 1997). However it should be noted that there are opinions considering the existing patent classification schemes are "never intended to provide conceptual delineations of technology areas, but instead identify inventions by function at very low levels of abstraction in order to serve as aids to prior art searching" (Allison et al., 2004).

As described above, PCA can be used to observe the focus of an entity's R&D activities up to the time of analysis or, if the entity's latest patents are gathered, of the entity's recent R&D activities. However, what is revealed by the latter is actually not the R&D activities happened around the time of analysis but a considerable amount of time in the past. To see this, the curve with diamond marks in Figure 2 depicts the distribution of U.S. utility patents issued in the year 2012 according to their application years. About three quarters of the 2012-issued utility patents are actually filed between 2007 and 2010. In other words, if a histogram similar to Figure 1 is derived from these 2012-issued patents, the revealed R&D focuses actually occur and disperse in a period of time quite in the past.

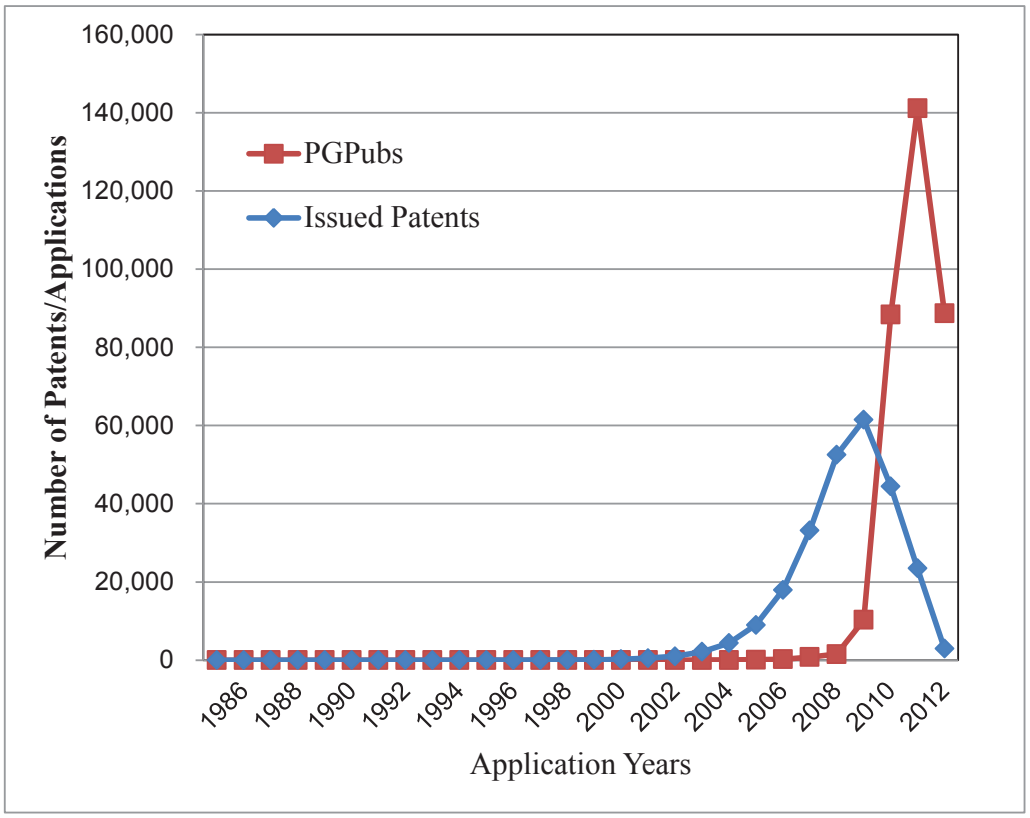


Figure 2. Distributions of 2012 issued patents and PGPubs based on application years.

The only possible way to reduce this time delay is to use the so-called *pre-grant publications* (PGPubs), instead of the issued patents. A patent application usually undergoes an early publication process before the patent is issued by the authority or before the patent application is given up by the applicant. Again taking Figure 2 as example, the curve with square marks depicts the distribution of U.S. PGPubs published in the year 2012 according to their application years. As illustrated most PGPubs are filed between 2010 and 2012, which are concentrated in a more limited period of time and in a more recent past.

The early publication process is a common practice for authorities across various nations and regions. For example, U.S. Patent Act (35 U.S.C. § 122(b)) specifies that, "each application for a patent shall be published … promptly after the expiration of a period of 18 months from the earliest filing date for which a benefit is sought under this title." There are indeed exceptions that an application is not early published if the application is (i) no longer pending; (ii) subject to a secrecy order; (iii) a provisional application; (iv) an application for a design patent; or (v) requested by the applicant. These exceptions are not common and, for utility patent applications, which are the most common type of patent applications, it is very possible that an issued utility patent is early published. According to our statistics, there are 253,580 utility patents issued in the year 2012 and 17,993 of them (7.1%) do not have corresponding PGPubs.

When a patent application is filed, the patent application is initially classified and classification symbols are assigned so as to route the patent application to an appropriate examiner team (USPTO, 2004). Then, after the patent application has undergone substantive examination, its examiner may alter the initial classification and assign different classification symbols (USPTO, 2005). As such a PGPub and its subsequently issued patent have their respective classification symbols and their classification symbols may not be identical.

PCAs usually utilize the issued patents, instead of PGPubs, most likely due to that the PGPubs have not undergone substantive examination, and their classification symbols may

not fully reflect their inventive contents. Yet PGPubs are better subjects for investigating the latest R&D focuses as they do not suffer lengthy pendency and strict screening by the examination process as reflected in Figure 2.

This study therefore tries to investigate the adequacy of using PGPub classification symbols for PCA. If the answer is yes, analysts can effectively reduce the time delay of their analytic observations to about 18 months, which is a significant improvement. On the other hand, even if the answer is no, analysts would know that PGPub classification symbols are not reliable, and they should avoid using them or at least be cautious about PCAs based on PGPub classification symbols.

Methodology

To investigate the adequacy of PGPub classification symbols for PCA, we collected U.S. utility patents issued in 2012 and their corresponding PGPubs for comparison. Utility patent is chosen because, for the three types of U.S. patents, utility patent is the most common and numerous one, design patents do not undergo the early publication process, and there are only a small number of plant patents. According to our statistics, there are only 868 plant patent applications filed each year between 1992 and 2011 on the average.

Each U.S. utility patent/PGPub is classified with three classification schemes: IPC, CPC, and USPC, and we choose the USPC symbols for comparison. This is because USPC is the default scheme for United States Patent and Trademark Office (USPTO) (USPTO, 2012), the IPC symbols are most likely machine-converted from the USPC symbols, and the CPC are not popular yet. Most importantly, USPC scheme does not have versions as it is updated every two months and the USPC symbols of all documents contained in USPTO databases are thoroughly and automatically re-classified accordingly (Wolter, 2012). In other words, when the USPC symbols of an issued patent are compared against those of its PGPub, whether the USPC symbols are of the same version is not an issue. One may question that USPC, as a domestic scheme, may not be representative. However, we believe that what this study observes from using U.S. patents and USPC could provide us at least some hint when dealing with patents of different countries and using different classification schemes.

Like all other classification schemes, USPC provides a hierarchical taxonomy of technical areas. Each USPC symbol contains a class code and a subclass code separated by "/." For example, a USPC symbol 623/2.1 has class code 623 and subclass code 2.1. The class code (e.g., 623) represents a highest level of non-overlapping technical area whereas the subclass code (e.g., 2.1) represents a lower level of technical area belonging to the one denoted by the class code. For subclass codes under the same class code, they may have hierarchical relationship among themselves. For example, 623/2.11 and 623/2.12 represent parallel technical areas but the two technical areas both belong to the technical area denoted by the symbol 623/2.1 (USPTO, 2012).

A U.S. utility patent/PGPub is assigned one or more USPC symbols. Among them, one and only one is expressed in boldface in the patent/PGPub documents. For issued patents, the official name for the bold-faced symbols is *original classification* symbols and, for the normal-faced symbols, *cross-reference classification* symbols by USPTO. As to PGPubs, the official name for the bold-faced ones is *primary classification* symbols and, for the normal-faced ones, *secondary classification* symbols. For simplicity's sake, we refer to the bold-faced symbols as the *main classification* symbols whereas the rest of the normal-faced symbols as the *auxiliary classification* symbols, whether or not they are from issued patents or PGPubs. The main classification covers the novel and non-obvious information contained in a patent/PGPub whereas the auxiliary classification covers other information considered to be valuable for searching (USPTO, 2012).

To determine whether PGPub classification symbols is adequate for PCA, we use the classification symbols assigned to the corresponding issued patents as reference as they are assigned by examiners after substantive examinations and therefore assumed to have better reflected the inventive contents of the patents.

Table 1 provides a number of examples where the sets of classification symbols assigned to three U.S. utility patents issued on 2015/02/10 and their PGPubs are listed side by side for comparison. As illustrated in Table 1, the two sets of classification symbols may not be identical, and the set assigned to the issued patent indeed seems to be more detailed than that assigned to the corresponding PGPub.

Table 1. The classification symbols assigned to three sample pairs of PGPubs/patents.

PGPub no./Patent no.	*PGPub symbols*	*Patent symbols*
20140289912/8,955,161	**850/18**	**850/1**; 250/339.11; 250/339.14; 73/105; 850/5; 850/50; 850/6
20120124680/8,955,160	**726/34**	**726/34**
20110252484/8,955,159	**726/32**	**726/32**; 380/201; 705/57; 726/27; 726/31; 726/33

There are quite some researches involving the measurement of similarity between nodes in a hierarchical taxonomy of concepts, which can be applied to classification symbols as well. For example, in one so-called edge-based approach, the similarity between two nodes is calculated based on the numbers of edges from the root of the hierarchical structure to the two nodes and to their nearest common ancestor node (Slimani, Yagahlane, & Mellouli, 2008). Similar edge-based approaches can be found in McNamee (2013). There are also so called node-based approaches, which capture a node's feature in the hierarchical structure as a vector and calculate a similarity measure based on the concept vectors of two nodes (cf. Liu, Bao, & Xu, 2012).

These studies do have their academic merit but cannot directly tell us whether PGPub classification symbols is reliable or not for PCA. We therefore adopt a different and practical treatment to the comparison of the classification symbols. First, we notice that existing commercial analytic systems/services conduct PCA using one of three simple approaches:

- PCA using Approach 1 counts only the class codes of the patent or PGPub main classification symbols so as to obtain a broad picture of the distribution of R&D activities;
- PCA using Approach 2 counts only the main classification symbols and ignores all auxiliary classification symbols of patents or PGPubs, considering that the main classification symbols are the most representative ones; and
- PCA using Approach 3 counts all patent or PGPub classification symbols with no distinction between main and auxiliary classification symbols, believing all classification symbols are equally important.

To demonstrate the three approaches, using the Patent Symbols column listed in Table 1 as example:

- Approach 1 counts the class codes 850 as being assigned once, 726 being assigned twice;
- Approach 2 counts each of the main classification symbols 850/1, 726/34, and 726/32 as being assigned once; and
- Approach 3 counts each of the 14 classification symbols as being assigned once.

Please note that, to the authors' knowledge, commercial analytic systems/services ignore the hierarchical relationship between classification symbols. For the above example, 850/6 is actually a technology area belonging to that of 850/5 but commercial analytic systems conducting PCA using Approach 3 treat 850/5 and 850/6 as denoting distinct technology areas probably for simplicity's sake.

Then, to see whether PCA using one of the above approaches on PGPubs classification symbols would deliver trustworthy result, we conduct three analyses as follows, each corresponding to one of the approaches above:

- Analysis 1 compares the main classification class codes of PGPubs to those of the corresponding issued patents.
- Analysis 2 compares the main classification symbols of PGPubs to those of the corresponding issued patents and calculates the consistency rate.
- Analysis 3 compares the sets of classification symbols of PGPubs to those of the corresponding issued patents.

Then all three analyses calculate the percentage of PGPubs having *identical* main classification class codes, main classification symbols, and sets of classification symbols to their corresponding issued patents. Since commercial analytic systems/services ignore the hierarchical relationship between classification symbols, our three analyses follow the same practice.

A 100% percentage indicates that PCA on PGPubs using one of the approaches would yield a result identical to that using their issued patents, meaning that using PGPubs can achieve reduced time delay with total accuracy. But a 0% percentage implies that PCA on PGPubs using one of the approaches delivers totally incorrect result. We therefore specifically refer to the percentage as *consistency rate* so as to avoid confusion with the general term *percentage*.

If statistically there is a very high consistency rate or similarity from the PGPubs, a histogram such as Figure 1 obtained from PGPubs using Approach 1, 2, or 3 would be very close to one from the corresponding subsequently issued patents. An analyst then can confidently utilize the PGPubs for PCA by Approach 1, 2, or 3 and achieve a reduced time delay.

To demonstrate the three analyses, again using the three sample pairs of PGPubs/patents listed in Table 1 as example:

- Analysis 1 shows that PCA using Approach 1 on PGPubs has a 100% consistency rate (i.e., all three pairs' PGPubs have identical main classification class codes to those of their issued patents);
- Analysis 2 shows that PCA using Approach 2 on PGPubs has a 66% consistency rate (i.e., except the first pair, the other two pairs' PGPubs have identical main classification symbols to those of their issued patents); and
- Analysis 3 shows that PCA suing Approach 3 on PGPubs has a 33% consistency rate (i.e., only the second pair's PGPub has an identical set of classification symbols to that of its issued patent).

For PCA using Approach 3, the simple consistency rate described above is too narrow to give us a complete picture. For example, even though the two sets of classification symbols from the third pair of patent/PGPub listed in Table 1 are different, the PGPub classification symbol {726/32} is actually a proper subset of the issued patent's classification symbols {726/32, 380/201, 705/57, 726/27, 726/31, 726/33} and therefore still captures a portion of the inventive content. The calculation of the consistency rate however ignores this condition.

Therefore in conducting Analysis 3, we divide the PGPub-patent pairs into 5 categories based on the relationships between their sets of classification symbols so as to gain more insight.

- Category 1: their sets of classification symbols are identical (i.e., $\{\text{PGPub}\} = \{\text{Patent}\}$).
- Category 2: their sets of classification symbols are entirely different (i.e., $\{\text{PGPub}\} \neq \{\text{Patent}\}$ and $\{\text{PGPub}\} \cap \{\text{Patent}\} = \emptyset$).
- Category 3: the PGPub's set of classification symbols is a proper subset of that of the corresponding patent (i.e.,$\{\text{PGPub}\} \neq \{\text{Patent}\}$ and $\{\text{PGPub}\} \subset \{\text{Patent}\}$).
- Category 4: the patent's set of classification symbols is a proper subset of that of the corresponding PGPub (i.e.,$\{\text{PGPub}\} \neq \{\text{Patent}\}$ and $\{\text{Patent}\} \subset \{\text{PGPub}\}$).

- Category 5: their sets of classification symbols are not entirely different, do not belong to each other, and have a non-empty intersection (i.e.,$\{\text{PGPub}\} \neq \{\text{Patent}\}$,$\{\text{PGPub}\} \not\subset \{\text{Patent}\}$, $\{\text{Patent}\} \not\subset \{\text{PGPub}\}$, and $\{\text{Patent}\} \cap \{\text{PGPub}\} \neq \emptyset$).

Then, for the patent/PGPub pairs belonging to each category, we calculate an average Jaccard Coefficient (Jaccard, 1901) as expressed in (1) where {PGPub} and {Patent} are the two sets of classification symbols assigned to the PGPub and the corresponding issued patent, respectively. Jaccard Coefficient, or Jaccard Index, or Jaccard Similarity Coefficient, was originally designed for comparing similarity between sample sets, and has already been applied in patent bibliometrics such as co-citation analysis (Small, 1973). Here we use it to capture the degree of discrepancy between {PGPub} and {Patent}.

$$J = \frac{|\{\text{PGPub}\} \cap \{\text{Patent}\}|}{|\{\text{PGPub}\} \cup \{\text{Patent}\}|} \quad (1)$$

Findings

We collected 253,580 utility patents issued in the year 2012 from USPTO database. After removing those having no corresponding PGPub, those having no classification symbol (e.g., these patents are withdrawn and withdrawn patents do not have patent classification symbols recorded in the USPTO database), and for unknown reason those having no main classification symbols, there are total 234,966 patents eligible for analysis. As mentioned in the previous section, USPC is updated every two months and all patents are re-classified accordingly. We collected the USPC symbols assigned to the 234,966 patents and their corresponding PGPubs under the USPC scheme up to 2013/10/31.

An initial statistics shows that the 234,966 patents have average 3.9 USPC symbols and their corresponding PGPubs have average 2.2 USPC symbols, and that 64.16% of the 234,966 patents have a greater number of USPC symbols than that of the corresponding PGPubs, indicating that issued patents seem do have more careful assignment of classification symbols than their PGPub counterparts. In some extreme cases, PGPub No. 2010/0316607 has the greatest number of USPC symbols (48) among all PGPubs whereas patent No. 8,179,540 has the greatest number of USPC symbols (65) among all patents. The latter is also the case having the greatest difference (63) between the issued patent and the corresponding PGPub.

Analysis 1

For each pair of the 234,966 PGPubs and corresponding issued patents, we compared the class code of the PGPub's main classification symbols against that of the corresponding issued patent, and we found that the consistency rate is 77.89%. That is, 183,024 out of the 234,966 pairs of PGPubs and patents have identical main classification class codes, and the remaining 51,942 pairs (22.11%) have difference main classification class codes. In other words, there is a 22.11% probability that a PGPub's main classification class code does not accurately reflect the inventive content of the corresponding patent.

Analysis 2

For each pair of the 234,966 PGPubs and corresponding patents, we compared the main classification symbol of the PGPub against that of the corresponding issued patent, and we found that the consistency rate drops to only 36.42%. That is, 85,584 out of the 234,966 pairs of PGPubs and patents have identical main classification symbols, and the rest 149,382 pairs (63.58%) have different main classification symbols. In other words, there is a very

significant 63.58% probability that a PGPub's main classification symbol does not accurately reflect the inventive content of the corresponding patent.

Analysis 3

For the 234,966 pairs of PGPubs and corresponding patents, we categorized them into 5 categories based on the relationships between their sets of classification symbols, and calculated the average Jaccard Coefficient for each category. The result is summarized in Table 2.

Table 2. Comparison result from Analysis 3.

Category	*Pairs*	*Percentage*	*Avg. Jaccard Coefficient*	*Std. Deviation*
1	14,958	6.37%	1	0
2	89,981	38.30%	0	0
3	63,057	26.84%	0.34	0.16
4	10,693	4.55%	0.45	0.15
5	56,277	23.95%	0.22	0.11

As illustrated, PGPubs in Category 1 are those having identical sets of classification symbols to their issued patents and their share (6.37%) among the 234,966 PGPubs is exactly the consistency rate of Analysis 3.

PGPubs in Category 2 are those having totally different sets of classification symbols from their issued patents and, for a PCA on these Category-2 PGPubs using Approach 3, the analytic result would be totally incorrect, but PGPubs of this category has the greatest share (about 38%) among all PGPubs.

PGPubs in Category 3 are those having sets of classification symbols being proper subsets to those of their issued patents, and cover about 27% of all PGPubs. For these Category-3 PGPubs, their classification symbols capture only 34% of the inventive content as reflected by their average Jaccard Coefficient. We can imagine that, for a PCA on Category-3 PGPubs using Approach 3, a histogram such as Fig. 1 would miss a significant amount of information.

Category 4 is a special case where PGPubs have sets of classification symbols that are proper supersets to those of the corresponding issued patents, and therefore covers the smallest share (less than 5%). For these Category-4 PGPubs, their classification symbols capture all inventive content but unfortunately provide on the average 55% (1-0.45) surplus and erroneous information. Again we can imagine that a histogram from PCA on Category-4 PGPubs using Approach 3 would contain too much noise.

Category 5 is a combination of Categories 3 and 4, meaning these 24% of the PGPubs have sets of classification symbols that not only miss significant amount of information but also provide significant amount of erroneous information, as reflected by the very limited average Jaccard Coefficient (0.22).

Conclusion

This study arises out of an attempt to use PGPub classification symbols for PCA so as to investigate an entity's latest R&D focuses with limited time delay. It is however speculated that the PGPub classification symbols are not carefully assigned and their adequacy for PCA has to be determined first.

We therefore gathered 234,966 pairs of issued patents and corresponding PGPubs, and compared their classification symbols in accordance with the three approaches that a commercial patent analytic system/service usually employ.

Assuming that the classification symbols of the corresponding issued patents better reflect the inventive contents of the patents and as such using them as reference, we find that, if the commercial patent analytic systems/services count the main classification symbols, or the entire sets of classification symbols of the PGPubs for PCA, only 36.42% of the PGPubs have identical main classification symbols, and only 6.37% of the PGPubs have identical sets of classification symbols to those of the corresponding issued patents. PCA using PGPubs as described can hardly be considered as reliable.

The best candidate for using PGPubs in PCA is the PGPubs' main classification class codes. We find that as high as 77.89% of the PGPubs have identical main classification class codes to those of the corresponding issued patents. The main classification class codes, however, represent the broadest technical areas and using them to investigate R&D focuses would provide only limited insight.

This study can be further carried out as follows. In order to make the main classification class codes even more useful for PCA, the consistency rate for each individual class can be determined. For some classes that have statistically very high consistency rate, PGPubs assigned with these class codes can be used for PCA with high confidence whereas, for classes of low consistency rate, an analyst should avoiding using them for PCA.

Additionally, one may be curious about why some class codes reveal higher consistency rates than the others. We speculate that, for some well-developed technical fields, the consistency rates of their class codes would be high as the classification of the related technology should be familiar to the examiners whereas for emerging technical fields, the consistency rates of their class codes would be low as the examiners may have different opinions on what the related technology should be classified. The investigation of this speculation is currently under way.

If both reduced time delay and better analytic insight are required, an analyst would require a better tool that can take the hierarchical relationship among classification symbols into consideration. If this kind of tool is available, we speculate that some specific technical areas may reveal a high consistency rate or similarity measure even for PCA using Approaches 2 and 3. The identification of these specific technical areas and how reliable the PGPub classification symbols are in these specific technical areas can be further investigated.

Acknowledgments

This study is funded by the Ministry of Science and Technology, Taiwan, R.O.C., under Grant No. MOST 103-2221-E-011-115.

References

Allison, J.R., Lemley, M.A., Moore, K.A., & Trunkey, R.D., (2004), Valuable Patents. Georgetown Law Journal, 92, 435–479.

Henderson, R.M., Jaffe, A., & Trajtenberg, M., (1997), University versus Corporate Patents: A Window on the Basicness of Invention. *Economics of Innovation and New Technology, 5*(1), 19–50.

Jaccard, P., (1901), Étude comparative de la distribution florale dans une portion des Alpes et des Jura. *Bulletin de la Société Vaudoise des Sciences Naturelles, 37*, 547–579.

Jaffe, A.B., (1986), Technological opportunity and spillovers of R&D: Evidence from firms' patents, profits and market value. *The American Economic Review, 76*(5), 984–1001.

Jaffe, A.B., (1989), Characterizing the "technological position" of firms, with application to quantifying technological opportunity and research spillovers. *Research Policy, 18*(2), 87–97.

Leydesdorff, L., (2008), Patent Classifications as Indicators of Intellectual Organization. *Journal of the American Society for Information Science and Technology, 59*(10), 1582–1597.

Lerner, J., (1994), The Importance of Patent Scope: An Empirical Analysis. *RAND Journal of Economics, 25*(2), 319–333.

Liu, H.Z., Bao, H., & Xu, D., (2012), Concept Vector for Similarity Measurement Based on Hierarchical Domain Structure. *Computing and Informatics, 30*(5), 881–900.

McNamee, R.C., (2013), Can't see the forest for the leaves: Similarity and distance measures for hierarchical taxonomies with a patent classification example. *Research Policy, 42*(4), 855–873.

OECD (1994), The Measurement of Scientific and Technological Activities Using Patent Data as Science and Technology Indicators: Patent Manual. Paris: OECD Publishing. DOI: 10.1787/9789264065574-en.

Slimani, T., Yagahlane, B.B., and Mellouli, K., (2008), A new similarity measure based on edge counting. *Proceedings of the World Academy of Science, engineering and Technology, 23*, 773–777.

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, 24*(4), 265–269.

USPTO (2004), Pre-Grant Publicaiton (PGPub) Global Concept of Operations. USPTO. Available on-line at http://www.uspto.gov/web/offices/dcom/olia/aipa/PGPubConOps.pdf.

USPTO (2005), Handbook of Classification. USPTO. Available on-line at http://www.uspto.gov/web/offices/opc/documents/handbook.pdf.

USPTO (2012), Overview of the U.S. Patent Classification System (USPC). USPTO. Available on-line at http://www.uspto.gov/patents/resources/classification/overview.pdf.

Wolter, B., (2012), It takes all kinds to make a world–Some thoughts on the use of classification in patent searching. *World Patent Information, 34*(1), 8–18.

The New Development Trend of Chinese-funded Banks and Internet Financial Enterprises from Patent Perspective

Zhao Qu, Shanshan Zhang and Kun Ding

qz_31@sina.cn, shann1027@163.com, dingk@dlut.edu.cn
School of Administration and Law, WISE Lab, Dalian University of Technology, Dalian, 116085
(People's Republic of China)

Abstract
Relying on the perfect integration of Internet technology, new business format and financial services, the Internet finance is developing at an unexpected speed, bringing impacts to Chinese-funded banks in the traditional business and emerging areas such as customization. Based on the preliminary study of the close contact between Chinese-funded banks and Internet financial enterprises as well as the necessity of patent protection, the paper proposes a comprehensive analytical framework and makes statistical comparison between 5 well-known Chinese-funded banks and Alibaba Group's patents from the perspective of annual trend, collaboration, application organizations, citation and other characteristics with data up to 2014 collected from Derwent Innovations Index(DII). It builds a Derwent Manual Code co-occurrence network with time coordinate by combining with visual tools and quantized the respective patent focuses of banks and Internet financial enterprises from the perspective of frequency and burst. After analysing the patents' contents, the paper discusses the mode of patent assignment. Finally, according to the status of patents, the paper concludes the strategic layout of domestic banks and Internet financial enterprise's intellectual property protection to predict the trend of further competition and alliance.

Conference Topic
Patent Analysis

Introduction
The data of British magazine "Banker" showed that in 2014, 13 Chinese banks ranked among the world's top 100 banks. Among them, Industrial and Commercial Bank of China ranked No.1 with the fund scale of 2,076.14 billion U.S. dollars, followed by China Construction Bank, Bank of China and other Chinese-funded banks, highlighting the fast growth and significant expansion of Chinese-funded banks. Nevertheless, the rates of return on assets of these banks were less than 3%, indicating that although the overall profit scale of China's banking ranked No.1 in the world, its profitability was not the case. With the slowdown of economic growth, substantial promotion of interest rate liberalization and further standardization of banking regulation, it is difficult for banks to maintain rising profit by relying on traditional channels. Like a huge dam, commercial banks store the saving deposits and collaborative deposits, but now there is a gap in the dam and the initiator is Internet finance. In the extensive penetration of Internet technology, traditional financial industry is undergoing dramatic changes: financial services have become the area competed by major institutions. Investors' "financial outlook" is corrected and the process of interest marketization has been promoted virtually (SOHO, 2014). The release of small and micro enterprises and individual consumer market's demand for loan is accelerated and the financing market presents a thriving prospect. With huge dividends of reform as well as the progress of big data and cloud computing technology, the Internet financial innovation is increasingly deepening. The rapid rise of Internet financial enterprises obliges Chinese – funded banks to face the continuous overlapping business, increasing demand for product service, competition and challenges brought by the application of innovative technologies.
In the new era, the competition between Chinese-funded banks and Internet financial giants does not only stay in the extent of business coverage, and more importantly, it is a rigid form

of innovation, which has been highly concerned by famous financial institutions, especially international banks, and produced historical and substantial effect on financial markets, services, products and management (Chen, 2006). Meanwhile, as an important link of financial products and intellectual properties, patents reflect the high degree of innovation of bank and Internet financial enterprises in service and product development. Meanwhile, in the period of patient protection, the banks exclusively enjoy the market of the innovative product, increase extra profits and safeguard fundamental interests. Events including the determination of the United States on the patentability criteria of bank business methods in 1998 or the patent bulk purchase of Alibaba Group before the listing in the United States in 2014 indicated that the field of financial patent protection has always been a focus of people. With the constant innovation of e-commerce and in-depth integration of Internet and mobile communication network, transaction platforms and payment means represented by e-banking, online banking and mobile banking will be bound to become the main form of future financial services. This control of the patents closely related to high-tech may become constitutor of financial market rules.

Theoretical basis and analytical framework

The slight decline of net interest margin posed no threat to large banks like ICBC, and the real blow came from the endogenous market force, the counterattack of Internet financial enterprises. For example, Ali Group's financial system has fundamentally broken the ice of the domestic credit loan by the "one-stop" service of customer absorption, credit assessment, loan review and issuance via e-business platform, providing more possibilities to the SME's problem of "difficult financing and expensive financing". In addition, Ali Group does not only involve in traditional fields of commercial banks including deposits and loans, financing, payment and settlement, but resulting in profound impact on commercial banking services and business philosophy. The formal establishment of Zhejiang E-business Bank ("Ali Bank") in 2014 intensified the potential threat to traditional banks. The strengthening of intellectual property protection strategy fired the first shoot of the competition between domestic banking industry and Internet financing; meanwhile, to defend the intellectual property disputes with foreign companies, especially under the circumstances of Ali's listing in the United States, Chinese companies will be exposed to a wider range of patent competition, so the enhancement of information sharing, innovative alliance building (Feng, 2013), and especially the optimization of patent protection become particularly important.

Overseas research on the relationship between Internet finance and banks was significantly earlier than China. Chou, et al, believed the in-depth integration of Internet and bank caused a revolutionary upheaval to the banking sector (Chou & Chou, 2000); Tsai, et al held the customers of Internet financial enterprises and traditional commercial banks varied in age, which was related to the degree of acceptance of innovative technologies and uncertain risk factors (Tsai, Huang & Lin, 2005). Meyer pointed out compared with commercial banks, P2P platform has lower operating costs and higher utilization of funds (Meyer, 2007); Ocean believed Internet financial enterprises provided more convenient credit business than bank process (Tess, 2013).

Chen believed the pressure of commercial banks caused by Internet finance should not be overlooked, forcing commercial banks to accelerate the pace of reform and strengthen customer customization (Chen, 2014); according to the status quo of competition between Internet financial enterprises and traditional commercial banks, Wang proposed four competitive strategies such as growth-orient strategy and aggressive strategy (Wang & Wang, 2014) by using the SWOT analysis; Gong thought the Internet financial model would not shake the traditional business model and earning way of commercial banks in a short term, and commercial banks should seek new development opportunities by using the Internet

(Gong, 2013). The above literature study involved the impact of Internet finance on traditional commercial banks as well as the business model based discussion on how commercial banks deal with Internet finance. However, its analysis of the relationship between commercial banks and Internet finance from the perspective of patent and technological innovation is still a blank area. This paper makes econometric analysis of the patents of Chinese banking industry and Internet financial giants, providing important reference basis for the development and improvement of the related patent protection system and patent strategy, the comprehensive analytical framework is proposed as shown in Figure 1.

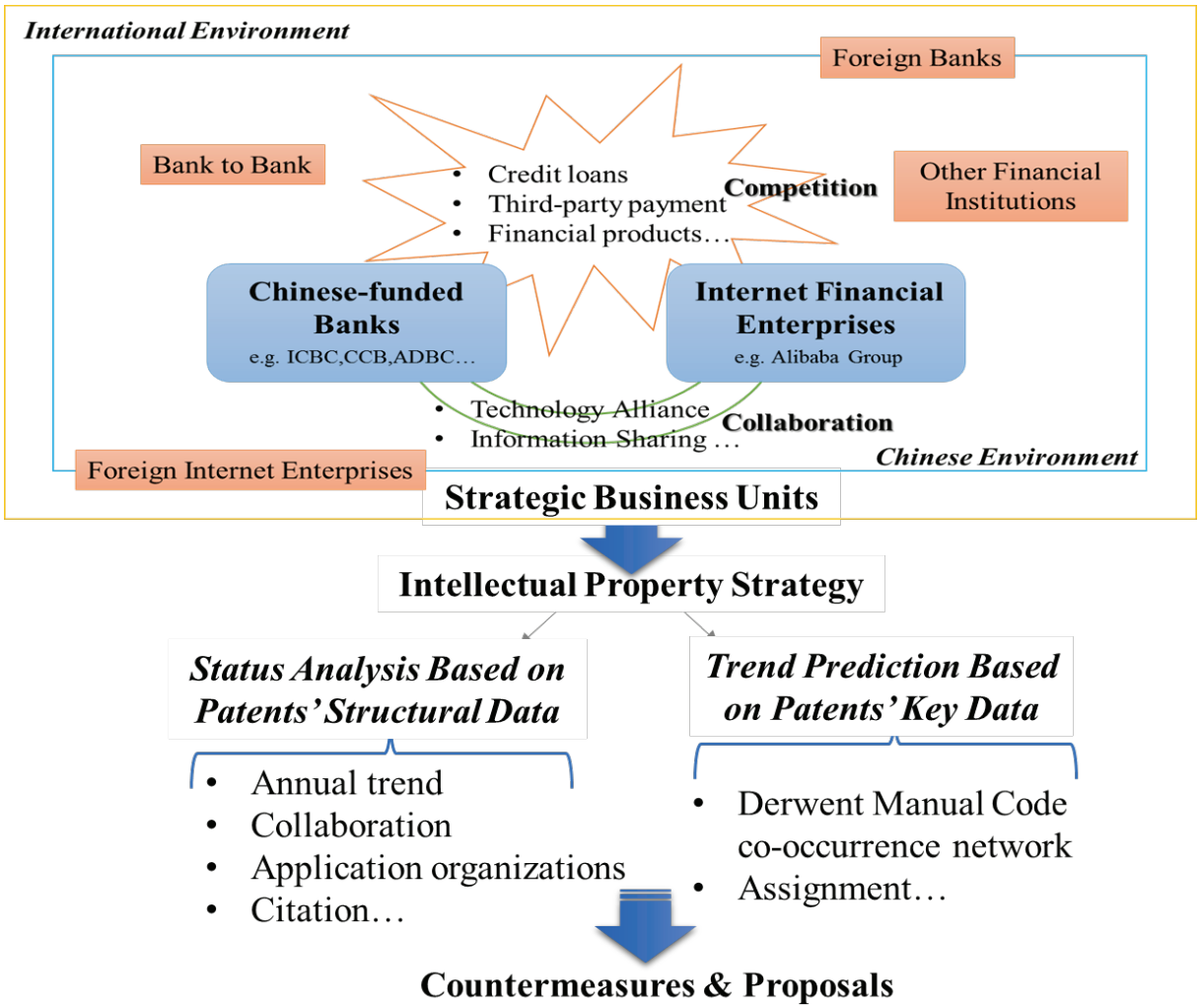


Figure 1. Patent analytical framework of Chinese-funded banks & Internet financial enterprises.

Data collection and analysis approach

The paper acquires the patents of the five representative Chinese-funded banks (ICBC, CCB, ADBC, BOC and BOCOM) and Alibaba Group Holding Limited on Jan.7, 2015 in DII by the way of Assignee Name and Assignee Code complex retrieval mode (Assignee Name and Assignee Code is connected by "OR" internally and by "AND" between two), the time span is from 1963 to 2014. After manual screening and exclusion, 917 Chinese bank patents and 1088 Ali patents are finally obtained.

The paper generalizes the patent development status and trend prediction of Chinese-funded banks and Internet financial enterprises by approaches of patent quantity statistical analysis

and patent content measurement in combination of visual tools, and proposes strategies and measures for the two sectors to improve patent protection, enhance technological innovation capacity, share information and build technology-business alliance if necessary, providing reference for the new development layout.

Results

Results of status analysis based on patents' structural data

Although the five Chinese-funded banks were built significantly earlier than Alibaba Group, they didn't occupy a striking advantage in the patent protection starting year, and lagged behind Ali in the total number of patents. In 2002, ICBC's patent of bank-card with dual account's processing device and method (PN: CN1397916-A) started the bank patent applications. Three years later, Alibaba carried out comprehensive patent protection and gradually exceeded the banks at an amazing growth. The annual patent application amount is shown in Figure 2.

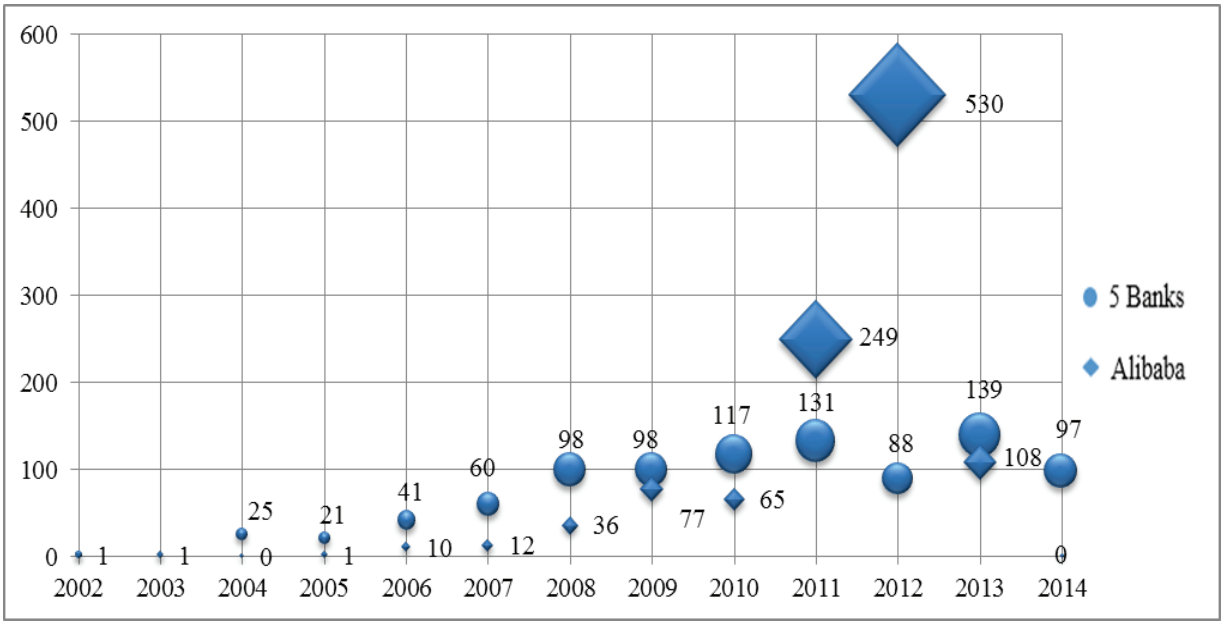


Figure 2. Annual trend of five Chinese-funded banks and Alibaba Group's patent quantity.

Figure 2 shows that the patent application amount of the selected banks has entered into fast growth since 2004. Though with slight fluctuation, but the overall situation is stable and the annual application number is relatively balanced. ICBC (549 patents) and CCB (253 patents) occupied a dominant position and led domestic banks to quickly engage in the patent development gradually integrating high-tech into the enterprise strategic level. In contrast, Ali Group's patent application was almost in exponential growth trend. The number of patent in 2012 was as high as 530, and the growth declined since 2013. The rapid deployment of domestic banks and financial enterprises was inseparable from the guidance of a series of policy documents including "National Intellectual Property Strategy" and also inseparable from the continuous expansion of Chinese enterprises and high-tech application.

By making statistics according to the patentee, we found all the 2005 patents were independently applied by banks and Ali Group. Few patents were produced via internal cooperation, and the branches concentrated in Zhejiang and Jiangsu. This phenomenon indicated that Chinese-funded banks and Internet financial enterprises didn't have close external relation in the patent activities, with a low degree of cooperation. To some extent, it

indicated that in the scope of finance, domestic enterprises have the relatively independent R&D team and were not positive enough in the flow and share of knowledge and information. If the external cooperation characterizes the degree of openness of proprietary technology, the geographical distribution of patent pending organizations is the indicator of measuring the corporate strategic deployment breadth. By the patent geological layout, we can learn and predict the key development areas of banks and Internet financial enterprises as well as the market distribution status of financial products and services (Luan, 2012). This paper makes analysis based on the connotation of the patent pending areas and organizations represented by the first two bits of code, we find only three patents of the Chinese-funded banks are applied in the non-Chinese mainland pending organizations, which are held by ICBC and distribute in WIPO, Taiwan and Russia. Although ICBC ranked No.1 in the world by a higher core capital and positively promoted international business strategy by means of organization application, mergers and acquisitions (till 2014, ICBC set up more than 330 overseas establishments in 41 countries and regions), its patent strategy failed to achieve the corresponding expansion (People, 2014). In contrast, Aliaba's patent has a wider geographical distribution; up to 71.7% (780pcs) of the patents were applied in organizations out of China. The average number of non-Chinese mainland patent application is 2.4 times (non-Chinese mainland application number/ non-Chinese mainland patent application number 1879/780), and the application of a number of patents has covered the range of over 6 organizations, and the pending mechanisms mainly distribute in Hong Kong, the United States and Europe (Table 1). Since the expansion of overseas business (since the establishment in 1998, Ali Group has set international headquarters in Hong Kong, offices in the United States, European and Japan), maintaining a highly consistent direction.

Table 1. Distribution of Ali's patent applications (outside of mainland China).

Region	*QTY*	*PCT(%)*	*Region*	*QTY*	*PCT(%)*
HK	631	33.58%	JP	186	9.90%
US	337	17.94%	KR	2	0.11%
WO	321	17.08%	SG	1	0.05%
EP	201	10.70%	AU	1	0.05%
TW	196	10.43%	DE	1	0.05%

Furthermore, the paper analyses status of two sections with patent citation data. These citations open up the possibility of tracing multiple linkages between inventions, inventors, scientists, firms, locations, etc. (Hall, Jaffe & Trajtenberg, 2001). 171 and 101 patents of Chinese banks and Ali Group were cited by other patents, respectively; patents with high citing frequency (top 5) were selected for analysis by combining with the cited patent information, and Table 2 is derived. Data showed that all the highly cited patents of Chinese banks were from ICBC, highlighting its outstanding R&D level among the peers.

Table 2. Highly cited patents of Ali and ICBC (Top 5).

ICBC		*Ali Group*	
PN/Freq. (cited patents)	AE/Freq. (citing patents)	PN/Freq. (cited patents)	AE/Freq. (citing patents)
CN1556449-A/19	BEIJ-Non-standard/10	CN101562543-A/7	GOOG-C/5
CN101183456-A/7	INCO-Non-standard/3	CN101662460-A	SALE-Non-standard/4
CN1588846-A/7	TNCT-C/3	CN101662460-A/6	IPCU-Non-standard/3
CN101119202-A/6	JIED-Non-standard/2	CN1835438-A/6	HUAW-C/2
CN101393671-A/5	SONG-Individual/2	CN101685516-A/5	TNCT-C/1

The patents of ICBC and Ali Group were mainly cited by enterprises, and a small number distributed in the patents held in the name of individuals and universities. Enterprises cited the patents of ICBC including categories of marketing, communications, telecommunications, network equipment, data security, authentication and other related categories, of which the citing frequency of BEIJING FEITIAN CHENGXIN SCI & TECHN CO (a world leading professional software protection and authentication of high-tech intelligence company), indicating the important of the authentication–related technology included in ICBC patents and also reflecting the close relation between the company products and ICBC business. Enterprises' citations of Ali Group involved customer consulting, Internet, software, communications (communications equipment), electronics, telecommunications, investing and financing, and the patent citers distributed in the United States and Japan. It is noteworthy that enterprises with similar business as Alibaba like Google, Tencent, are also among the citing group, showing Ali's patent technology is playing a guiding role in the Internet industry. In addition, Beijing Institute of Technology and Taiyuan University of Technology cited the patent of Ali and ICBC once, respectively.

Results of trend prediction based on patents' key data

Compared to other classification system, Derwent manual code (MC) outlines more detailed indexing information in retrieval of patent's theme and core content based on the uses and applications of an invention, rather than just a straight forward description of what the invention is (Stembridge, 1999).

Table 3. High frequency Derwent Manual Codes (Top 10).

Alibaba freq			*Five Chinese-funded banks*		
Freq	MC	Content	Freq	MC	Content
310	T01-J05B4P	Database applications	175	T01-J05A1	Financial
230	T01-N01D3	From remote site or server	140	T01-N01A1	Eft/banking
184	T01-S03	Claimed software products	139	T01-N01D3	From remote site or server
172	T01-N02A3C	Servers	134	T01-J05B4P	Database applications
154	T01-N03A2	Search engines and searching	81	T05-L03C1	General control system
126	T01-J05B3	Search and retrieval	75	T01-N02A3C	Servers
123	T01-N01D2	Document transfer	69	T01-D01	Data encryption and decryption
77	W01-A07G1	Transmission control procedure	67	T01-N01A	Financial/business
74	T01-N01A	Financial/business	59	T01-N01D2	Document transfer
65	T01-N02A2C	Client/server system	57	T01-N02B2B	System and fault monitoring

Further, we transforms the bibliographic data of all the 2005 patents into WoS logging data and introduced into the CiteSpace, and set the analysis interval as 1 year, then drawing the maps (Figure 3 and Figure 4). By depicting the association and combination between the MCs, it can analyse the correlation between patents and even technologies, and can also facsimile the internal technology composition and structure (Shen, Gao & Teng, 2012). Timeline visualization provides a directly temporal overview of technologies, columns are time periods of co-occurrence of technologies and rows are clusters (Gong, Jiang, Yang& Wei, 2011). The dynamically changing course of banks and Internet finance patent technologies can be revealed by combining with the attribute changes in timeline axis. Moreover, the development trend can be predicted through their restive business characteristics. The top 10 high-frequency manual codes of Chinese-funded banks and Ali Group (Table 3) were intercepted respectively to explore the hot fields.

It can be seen from the analysis that the technical research of both subjects was carried out by centring the category of "T01", showing the Chinese banks and Internet financial enterprises are very concerned about the application of digital computer in financial services. A series of patent activities were conducted by combining with the research of "database applications" and "application originating from remote sites or remote servers". It is noteworthy that in the distribution of the top 10 high-frequency bank patents, Internet financial patents showed a high degree of overlap in some technical contents. In addition to "database applications" and "remote service", "document transfer" and "Financial/business" were also included in the key content of their patent developments. In contrast, the patents of banks are more inclined to the study of financial, banking, system monitoring and related technology; Ali Group makes innovation and protection based on the contents of search engine and software.

As the largest cluster in the bank MC network, "bank background" demonstrated the general picture of banking business featuring electronic funds transfer point of sale equipment, currency handling systems, smart media and the Internet and information transfer, which occupied the central position in the entire time chain.

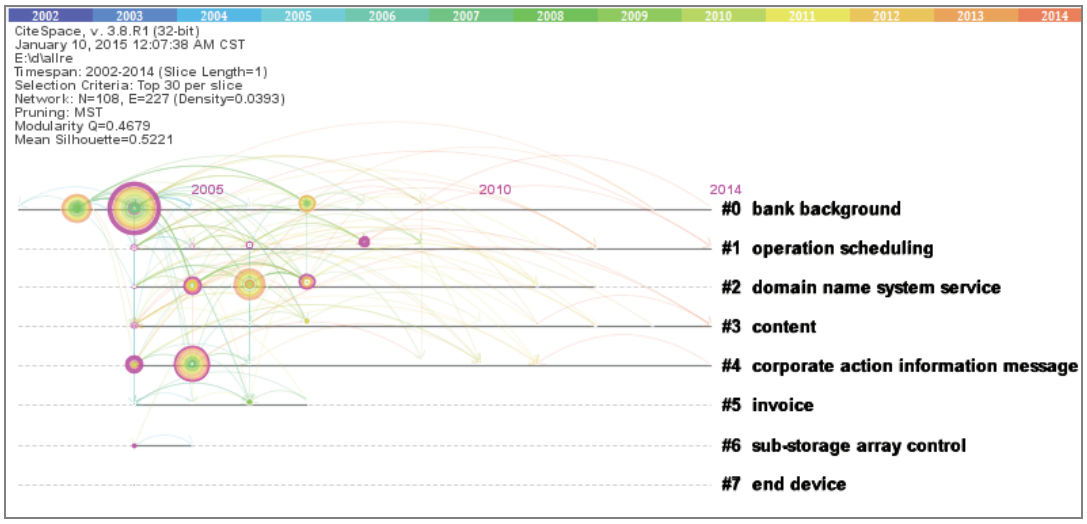


Figure 3. Five banks' Derwent Manual Code co-occurrence network (Timeline view).

Figure 4. Alibaba Group's Derwent Manual Code co-occurrence network (Timeline view).

In Ali's network, the cluster "outgoing message" constituted by the close connection of digital information transmission, Internet and messaging, data processing systems and process control comprehensively summarized the business flow carried out by Ali Group based on Internet data. Second, the cluster "central account registration system" composed by audio / video record and Internet-based information processing and transfer, and nine clusters including data and communications. The overall technology relevance and research contents are similar to these shown in the MC of Chinese bank patents, but more emphasis was made on the application of Internet in business.

On this basis, codes with high frequency change rate with the time sequence (burst term, Table 4&5) further determined the technology frontier and development trend of Chinese banks and Ali (Huang, Wang &Wang, 2014).

Table 4. Bursts of Banks' Derwent Manual Codes

Burst	*MC*	*year*	*Content*
5.77	T05-L03	2002	Cash dispensing and depositing machines
6.18	T05-L02	2003	Electronic funds transfer
5.21	T01-N01A1	2003	Eft/banking
3.06	T01-N01A2A	2004	E-shop, e-auction, e-mall, and e-services
2.94	T01-J05A1	2004	Financial
2.93	T05-L01D	2004	Data transfer and network aspects
2.76	T01-J12C	2004	Security
2.76	T01-J05B4P	2005	Database applications
5.7	T01-F05	2006	Arrangements for executing specific programs and system management software
4.93	T01-N01D	2006	Data transfer
3.53	T01-J05A2	2006	Administration and management tools
4.08	W01-A07G1	2011	Transmission control procedure
2.99	W01-A06C4	2011	Radio link
2.68	T01-N03A2	2011	Search engines and searching
3.04	T04-K03B	2012	Rfid/transponder

Table 5. Bursts of Ali' Derwent Manual Codes

Burst	*MC*	*year*	*Content*
5.12	T01-N01A1	2005	Eft/banking
3.14	T01-N02A3C	2006	Servers
5.02	T01-E01A	2007	Sorting
4.48	T01-S03	2007	Claimed software products
2.86	T01-J16C3	2007	Natural and pictorial language processing
4.58	T01-M02	2008	Multiprocessor systems
6.29	T01-E01	2009	Sorting, selecting, merging or comparing data
4.63	T01-J20C	2011	Software test, verification, debug, optimization
2.73	W01-A06E	2013	Network control and software

The patented technology burst of Chinese banks are more evenly dispersed in 2002~2012, following the development course of bank reserves appliances → electronic funds / bank → online business and data processing → database applications → specific project management and data transfer → search engine, control → wireless communications, showing the trend of gradual evolution from traditional banking to Internet financial sector. Since 2005, Ali's patent started from e-funds/e-bank technologies, and then underwent a series of technology evolution of data processing from server, data sorting, and software to graphic language processing, which is currently in the data processing optimization and study of Internet control technology. Although the related technologies of e-transaction technology appeared earlier in the patent of Chinese banks, but Ali Group is more sustainable in the ongoing online transactions, which continues to carry out the research based on big data and gradually establish technology chain in the field of Internet finance.

Technological evolution is the exploration on the development route and trend of bank and Internet financial enterprises based on patent, and the conclusion of patent assignment information can provide references to the patent development mode of the two. In 2014, Alibaba Group made IPO financing amounted to 25 billion U.S. dollars, which was the largest IPO. The United States is a country with frequent patent disputes, to avoid the patent infringement issues encountered by Facebook or Twitter in IPO, Ali Group has made significant patent deployment in the U.S. since 2013, where a lot of patents have been reserved. Till the retrieval date of this paper, 399 U.S. patent family cases were found and more than 50 have been authorized (Chinaip, 2014). In addition to independent application, Alibaba purchased 21 patents from IBM in 2013, and one of which was for Amazon, the largest U.S. e-commerce platform, and also prepared for coping with the patent competition and litigation. We made inquiry of the operating data of Ali Group and five Chinese banks in Chinese patent database and found that Ali Group started to purchase the patents of other organizations since 2012 onwards, but only limited to the category of invention patents. Patent seller expanded from domestic organizations to international institutions, such as Shanghai Yiren Information Technology Co., Ltd. and IBM; in addition to enterprises, Ali also purchased patents from Chinese Academy of Science Institute of Computing Technology; the change of some patent was caused by the changes of the corporate nature, such as Alibaba to Alibaba Group Holding Limited. The aforementioned technical fields of patent change included electric digital data processing, transmission of digital information, arrangements of circuit components or wiring on supporting structure and coin-freed or like apparatus. However, the patent purchased by Chinese banks included patent, utility models and appearance design, and the patents with internal change were almost 1/2 of the total patent transfer amount. These patents mainly came from the bank branches and individuals, and only CCB had one patent purchase from enterprise (Shandong Confucian Culture Communication Co., Ltd.), and the technical fields of patent change mainly involved the bank cards, security cards, teller settings and other contents, no transactions concerning goods and services of bank financial commodities and services were made.

Discussion and conclusions

General comments

In a long term in the past, Chinese banks made huge profits by relying on monopoly advantages and policy bonus, and occupied the position on the top of financial ecology. However, the single channel and curing product business model can no longer work. In China, the rapid development trend of Internet finance represented by Alibaba does not only occupy a significant share in domestic financial sector, but also causes widespread concern in the overseas business expansion. Traditional profit making channels of banks have been

hindered in a variety of aspects, including the competition of domestic and overseas banking industries and the pressure caused by the enhancement of overlap ratio with Internet finance business. With the development of commodities and services based on big data, Internet financial enterprises are inseparable from the application of technology. In the new situation, it faces the transfer from purely financial products to technical competition; whether banks or Internet financial enterprises, technology innovation and application have been upgraded to a new strategic plan.

By the comparison of patents of 5 Chinese banks and Alibaba Group Holdings Limited, we found that the patent activities of Chinese banks started late, with limited number, especially in key business areas like e-commerce. Most of the bank patents were independently applied in China, and their overseas IPR protection does not match their development of business, which may become a potential hazard for patent disputes arising from overseas promotion of financial products and services. Although the banks have higher patent citing frequency, the citing parties are mostly in China and the all the highly cited patents are held by ICBC. In contrast, Ali Group has achieved rapid progress of patent activities, with advantages in the total number, patent geographical distribution and the composition of citing groups. However, like banks, Ali Group also has low degree of external cooperation, indicating their closure and limitations in patent research and development. We can learn from MC co-occurrence network that banks and Internet financial enterprises have relatively concentrated technology, which were the patent R&D centred by computer and showed a high degree of overlapping in database use, financial/commercial and remote control, etc. The patent contents of Chinese patents tend to the research of digital communication, hardware equipment and banking business operation, whereas Alibaba pays more attention to search engine and software-related innovation and protection. From 2002 to 2014, bank patent technology showed the shift from bank reserves appliance to e-funds/banking, online services and data processing. Currently, it is in the stage of network and wireless communications, whereas the research of Alibaba has undergone a series of technology evolutions from e-funds/e-banking, data processing from server, data processing, software to graphic language processing. Patent assignment data showed that independently developed ones are still the main source of banks and Internet financial enterprises' patents, while the patent purchase of Internet financial enterprises are quietly rising, and may form a new patent development mode of "independent R&D and purchase".

Countermeasures & Proposals

Based on the abovementioned patent status and future development direction of banks and Internet financial enterprises, China's banking industry shall attach important to the development, protection, management and utilization of bank patents at all levels. Moreover, it is essential to set up product and service technology early warning, make technical prediction and selection in fields with priority. At the same time, cooperation with high-tech industries represented by information technology shall be emphasized to improve the patent technical quality. At the same time, on the basis of full study of international regulations and overseas local laws and regulations, Chinese banks shall learn from Alibaba's international patent strategies to increase the overseas patent application quantity, expand market share and gain competitive advantages. After the listing in the United States, as the leader of Internet financial industry, Alibaba shall not only strengthen the risk control effort, promote the innovation of financial products and services and customer participation as well, but shall accelerate the deployment of intellectual property, take the mode of simultaneous patent purchase and independent R&D, to avoid patent disputes with overseas companies and win market opportunities by appropriate use of patents. In addition to strengthening their competitive advantages, banks and Internet financial enterprises shall strengthen cooperation

to make best use of the advantages and bypass the disadvantages, so as to form a new finance-technology alliance. Banks can use the network resources, information data and cloud computing of Internet financial enterprises to play their professional administration, thus introducing customers to the professional advantages via network channel. Likewise, by relying on the financial background of banks, Internet financial enterprises shall set up long-term, stable relationship with mutual trust to expand the scope of commercial exchanges, strengthen financial risk management and control, thereby providing a cooperation and win-win opportunity to both parties.

Further research

In the process of researching the status quo and future trend of Chinese-funded banks and Internet financial enterprises, this paper only took into account of their competition and cooperation. In fact, we can learn from the framework of this paper that factors affecting the development of them are multifaceted and complex. Hence, in the following study, the author will put overseas companies into the comparison to explain the development situation of banks and Internet financial enterprises in detail.

Acknowledgments

We would like to thank the anonymous reviewers for their important comments. Thanks to Dr. Liu and Dr. Tang for his assistance in data cleaning.

References

Chen, J, M. (2006). The influence of science and technology revolution on the development of financial sector - our country uses technological innovation to promote financial innovation. *Journal of Dialectics of Nature*, 27(6), 99-100.

Chen, W. Y. (2014). What does commercial bank learn from Internet financial business? *Management and Administration*, *11*, 014.

China Intellectual Property (2014). Alibaba purchased patents to deal with patent litigation risk. *China Intellectual Property Platform*. Retrieved January 10, 2014 from: http://www.chinaipmagazine.com/news-show.asp?id=12219.

Chou, D. C., & Chou, A. Y. (2000). A guide to the Internet revolution in banking. *Information Systems Management*, *17*(2), 51-57.

Feng J.J. (2013). Research on competitive strategy of commercial bank under the background of the Internet finance. *Contemporary Finance*, *4*, 14-16.

Gong, X., Jiang, L., Yang, H., & Wei, F. (2011). Mapping Intellectual Structure: A Co-citation Analysis of Food Safety in CiteSpace II. Gene, 412.

Gong, X. L. (2013). The influence of Internet financial mode on traditional banking. *South China Finance*, *5*, 86-88.

Hall, B. H., Jaffe, A. B., & Trajtenberg, M. (2001). The NBER patent citation data file: Lessons, insights and methodological tools (No. w8498). National Bureau of Economic Research.

Huang L.C., Wang K., & Wang K.K. (2014). Technology Hot Spots and Fronts of Household Air Conditioner: Identification and Trend Analysis Based on CiteSpace. *Journal of Intelligence, 33*(2),

Luan C.J. (2012). Emperical Study on the Measuring Indicators of Generic Technology of Emerging Industries of Strategic Importance. *Forum on Science and Technology in China*, *6*, 73-77.

Meyer, T., Heng, S., Kaiser, S., & Walter, N. (2007). Online P2P lending nibbles at banks' loan business. *Deutsche Bank Research*.

People. (2014). Industrial and commercial bank branch opened in London, and has been set up more than 330 overseas agencies. *People.cn-Bank channel*. Retrieved December 30, 2014 from: http://finance.people.com.cn/money/n/2014/1202/c218900-26132904.html.

Shen J., Gao J., & Teng L. (2012). Derwent Manual Code Co-Occurrence: A Practical Method in Patent Map. *Science of Science and Management of S. & T.*, *33*(1), 12-16.

SOHU. (2014). Report on Top 16 Chinese Internet Financial Enterprises. *SOHO Media platform*. Retrieved December 10, 2014 from: http://stock.sohu.com/20140821/n403633305.shtml.

Stembridge, B. (1999). International patent classification in Derwent databases. *World Patent Information, 21*(3), 169-177.

Tess Ocean. (2013) Online Personal Loans: Access Easy Finance At Cheap Interest Rates By. Retrieved December 10, 2014 from: http://www.Internetmonetary.com.

Tsai, H. T., Huang, L., & Lin, C. G. (2005). Emerging e-commerce development model for Taiwanese travel agencies. *Tourism Management*, *26*(5), 787-796.

Wang, R. C. & Wang, J. Z. (2014). SWOT analysis of the commercial banks to deal with the Internet financial enterprise competition. *Small and medium-sized enterprise management and technology*, *8*, 62-63.

Who Files Provisional Applications in the United States?

Chi-Tung Chen[1] and Dar-Zen Chen[2]

[1] *d94522022@ntu.edu.tw*
Department of Mechanical Engineering, National Taiwan University, Taipei, Taiwan

[2] *Corresponding Author: dzchen@ntu.edu.tw*
Department of Mechanical Engineering and Institute of Industrial Engineering, National Taiwan University, Taipei, Taiwan

Abstract
This paper employed the US Patent Application Database to find out who files provisional applications in the United States. Preference rates, use rates, and provisional application to non-provisional application rates were used to evaluate the filing behaviour of provisional applications with respect to non-provisional applications. Factors weighing toward filing provisional applications include filing date sensitivity, patent term sensitivity, and necessity of promoting. Factors weighing against filing provisional applications include cost sensitivity and English abilities. These factors were discussed in order to explain the filing behaviour of provisional applications with respect to non-provisional applications. Applicants form English speaking countries are more likely to file provisional applications than applicants from other countries. We reasoned that the English ability of applicants might be the cause for such a result. Applicants from the fields of Computers and Communications and Drugs and Medical are more likely to file provisional applications than applicants from other fields. We reasoned that patent term sensitivity and filing date sensitivity might be the cause for such a result.

Conference Topic
Patent Analysis

Background and purpose
A provisional application for patent (hereafter referred to as 'provisional application') is a US national application filed in the United States Patent and Trademark Office (USPTO) that has been offered to applicants since June 8, 1995 and was designed to provide a lower-cost first patent filing in the United States. A provisional application is not required to have a formal patent claim or an oath or declaration. Provisional applications also should not include any information disclosure (prior art) statement since provisional applications are not examined. A provisional application provides the means to establish an early effective filing date in a later filed non-provisional patent application (hereafter referred to as 'non-provisional application'). It also allows the term "Patent Pending" to be applied in connection with the description of the invention. A provisional application has a pendency lasting 12 months from the date the provisional application is filed. The 12-month pendency period cannot be extended. Therefore, an applicant who files a provisional application must file a corresponding non-provisional application for patent during the 12-month pendency period of the provisional application in order to benefit from the earlier filing of the provisional application. By filing a provisional application first, and then filing a corresponding non-provisional application that references the provisional application within the 12-month provisional application pendency period, a patent term endpoint may be extended by as much as 12 months. (USPTO, 2014).

Although the provisional application filing approach has been offered to applicants for almost two decades, the USPTO does not make its database of provisional applications publicly available other than the individual files in Patent Application Information Retrieval (PAIR). Therefore, it is still difficult to answer the following two crucial questions: (1) Who files provisional applications in the United States? (2) Why do applicants file provisional applications in the United States?

Dennis Crouch (2008) studied approximately 15,000 utility patents issued in April and May 2008 and found out that only 21% of issued patents claiming priority from a provisional application, only 5% of the patents that associated with a provisional application were assigned to international applicants while 30% of the patents that associated with a provisional application were assigned to a U.S. applicant, Israel and Canada filed the highest proportion of provisional parent claims, only 2% of the Japanese & Korean patents included provisional parent claims, new drug inventions have the highest rate of association with a provisional application, and patents on electrical and electronic applications had the lowest rate of provisional filing. Dennis Crouch provided a rough first look of provisional application filings in the United States, but the dataset used by Dennis Crouch was rather small and time-limited (approximately 15,000 utility patents issued in April and May 2008). Therefore, it seems that the dataset used by Dennis Crouch was not sufficiently large to guarantee the results; and moreover, Dennis Crouch provides the results but lacked to explain the results.

The purpose of this paper is to address the two questions identified with sufficient dataset and detailed analyses to guarantee the results and to fully understand the filing behaviour of applicants. First, we employ the US Patent Application Database for 2005-2013 to find out who files provisional applications by checking the provisional application filings in different countries of origins, technological categories, assignee types, and assignees. Second, we explain why applicants file provisional applications in the US According to the USPTO, most obvious advantages of filing a provisional application are: (1) obtaining an effective filing date with a lower cost and an easily prepared application; (2) extending the statutory patent term up to one year; and (3) the ability to use the term "patent pending" (USPTO, 2014). Therefore, we assume that the following factors are weighing toward filing provisional applications: (1) filing date sensitivity; (2) patent term sensitivity; and (3) the necessity of promoting. Although the provisional application is designed to provide a lower-cost first patent filing in the US, an applicant still needs to spend extra money to file a corresponding non-provisional application in order to obtain a patent. In addition, although the provisional application was supposed to be an easily prepared application as it may be filed in a foreign language, an applicant still requires the English ability to prosecute the provisional application. Therefore, we assume that the following factors are weighing against filing provisional applications: (1) cost sensitivity; and (2) the English ability of applicants.

Trends in filing provisional applications

Since the database of provisional applications is not published, the filing numbers of the provisional applications can only be obtained from annual fiscal reports by the USPTO. Moreover, since the USPTO has never made publicly available the provisional applications that are not relied on for claiming priority by non-provisional applications, we employed the USPTO Patent Application Database to find out the number of provisional applications that have been claimed for priority by at least one non-provisional application.

Figure 1 shows the trends in filing provisional applications. The black bars represent the number of utility applications (non-provisional applications) filed each year from 2005 to 2013; the hatched bars represent the number of provisional applications filed each year from 2005 to 2013; and the grey bars represent the number of provisional applications filed each year from 2005 to 2013 that are relied on as priority documents in non-provisional applications. Please note that the USPTO only reported the number of provisional applications by fiscal year. So in Figure 1, the hatched bars were calculated by the fiscal year (October 1 to September 30), not by the calendar year (1 January to 31 December).

As shown in Figure 1, from 2005 to 2013, over 4.29 million non-provisional applications and over 1.27 million provisional applications have been filed. Among the 1.27 million

provisional applications, over 0.71 million provisional applications have been converted to non-provisional applications. It can be inferred that both non-provisional application filings and provisional application filings continued to rise, with over 570,000 and 170,000 filed in 2013. There was a drop in each of the non-provisional application filings and the provisional application filings in 2009. A possible explanation for such a drop could be attributed to the financial crisis of 2008.

Figure 1 also shows the provisional applications that have been relied on for claiming priority by non-provisional applications. It is observed that the number of provisional applications that have been relied on for claiming priority by non-provisional applications is growing. Although the provisional applications continued to be more popular, applicants have abandoned more of the provisional applications without relying upon them for claiming priority. The difference between each pair of the hatched bar and the grey bar is the number of provisional applications abandoned without being used as priority documents each year.

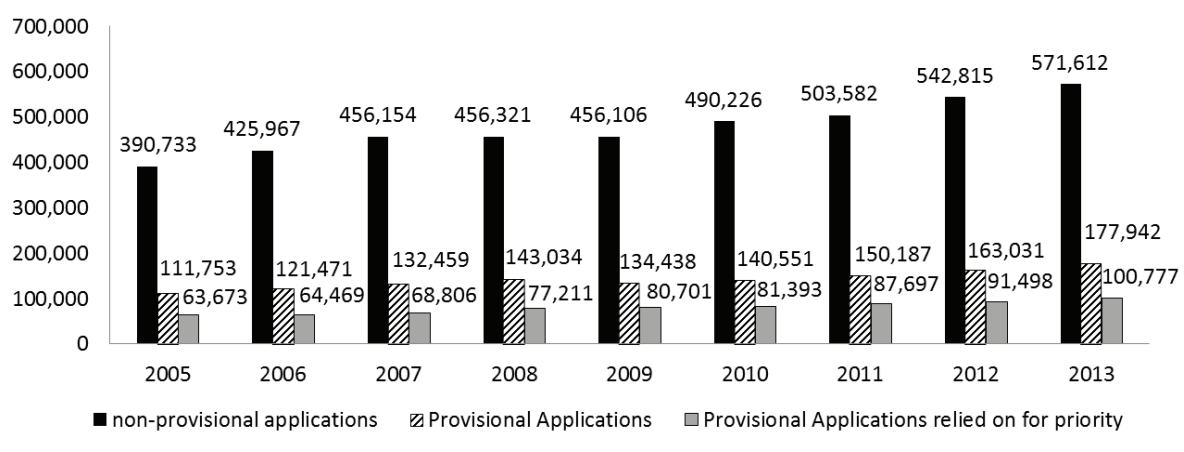


Figure 1. Non-provisional applications, provisional applications, and provisional applications relied on for priority filed each year for 2005-2013.

Rates of provisional applications/non-provisional applications

Rates of provisional applications/non-provisional applications (hereafter referred to as preference rates) show the preference of applicants in filing provisional applications with respect to non-provisional applications. The preference rate represents the percentage of a provisional application being filed in proportion with a non-provisional application in deciding filing patent applications in the United States. In Figure 2, the dotted line shows the preference rate of all provisional applications filed each year from 2005 to2013. It is clear that the preference rate remained steady during the period, except for 2009-2010, and the preference rate continued to slightly rise to 31.13 % in 2013.

Rates of provisional applications relied on for priority /provisional applications

As mentioned above, a provisional application has a pendency lasting 12 months from the date the provisional application is filed. An applicant who files a provisional application must file a corresponding non-provisional application for patent during the 12-month pendency period of the provisional application in order to benefit from the earlier filing of the provisional application (USPTO, 2014); otherwise, the provisional application will be automatically abandoned. Therefore, it is interesting to find out the use rate of the provisional applications that have been used for claiming priority by non-provisional applications (hereafter referred to as use rate). The use rate represents the usage of provisional applications. The result is shown in Figure 2, where the first solid line represents the use rate of all provisional applications filed each year from 2005 to 2013. As shown in Figure 2, the use rate

of provisional applications was located between about 52% and about 60% in 2005-2013, that is, about 40% to about 48% of the provisional applications were abandoned without being converted to non-provisional applications each year during 2005 and 2013.

Rates of provisional applications relied on for priority/non-provisional applications

Rates of provisional applications relied on for priority/non-provisional applications (hereafter referred to PA to NPA rate) show both the filing preference and the usage of provisional applications. The PA to NPA rate can be calculated by the preference rate times the use rate. Since the USPTO has never mad publicly available the provisional applications that are not relied on for claiming priority by non-provisional applications, the PA to NPA rate became the only practical rate for evaluating the provisional application filings with respect to non-provisional application filings in different countries of origins, technological categories, and assignees. As shown in Figure 2, the second solid line represents the PA to NPA rate of all the provisional applications filed each year between 2005 and 2013. It can be seen that the PA to NPA rate remained steady during the period, except for 2009-2010, and it continued to slightly rise to 17.63% in 2013. In other words, approximately one in six non-provisional applications was expected to claim priority upon a provisional application.

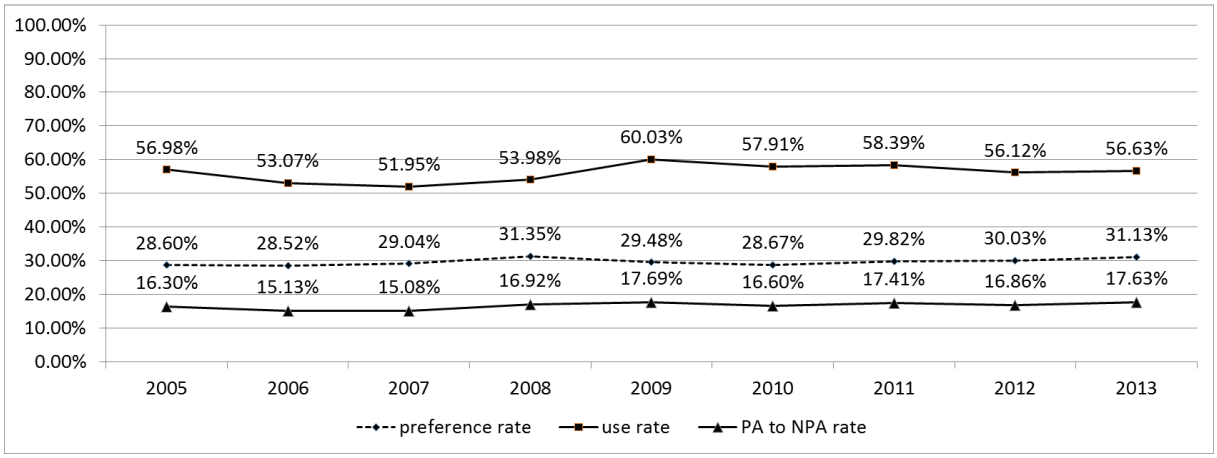


Figure 2. Preference rate, use rate and PA to NPA rate each year from 2005-2013.

Provisional applications by different countries of origins

The date of the filing of the provisional patent application can also be used as the foreign priority date for applications filed in countries other than the United States. Therefore, the need is identified for a foreign applicant to file a patent application as a provisional application in the United States first, and then to claim the priority of the provisional application to file a regular patent application in the United States as well as in the countries other than the United States.

Table 1 shows the ranking of the top 10 countries of origins where applicants filed provisional applications and non-provisional applications in the US in 2005-2013. During this period, the top 10 countries were: United States of America (US), Canada (CA), Germany (DE), Japan (JP), Israel (IL), Netherlands (NL), Korea (KR), Taiwan (TW), France (FR), and Switzerland (CH). It can be seen in Table 1 that the ranking of provisional applications and that of non-provisional applications varied for some countries. For example, JP was ranked second in non-provisional applications but fourth in provisional applications; KR was ranked fourth in non-provisional applications but seventh in provisional applications; TW was ranked fifth in non-provisional applications but eighth in provisional applications; FR was ranked sixth in non-provisional applications but ninth in provisional applications; and CN (China) was

ranked seventh in non-provisional applications but was not ranked in the top ten in provisional applications. It can be concluded that applicants in JP, KR, TW, FR and CN prefer filing their first applications in the United States as regular non-provisional applications rather than provisional applications. On the contrary, applicants in the US, CA and IL very much prefer filing their first applications in the US as provisional applications.

Table 1. Ranking of the top 10 countries of origins where applicants filed provisional applications and non-provisional applications in the US in 2005-2013.

ranking	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*
provisional applications	US	CA	DE	JP	IL	NL	KR	TW	FR	CH
non-provisional applications	US	JP	DE	KR	TW	FR	CN	NL	CA	GB

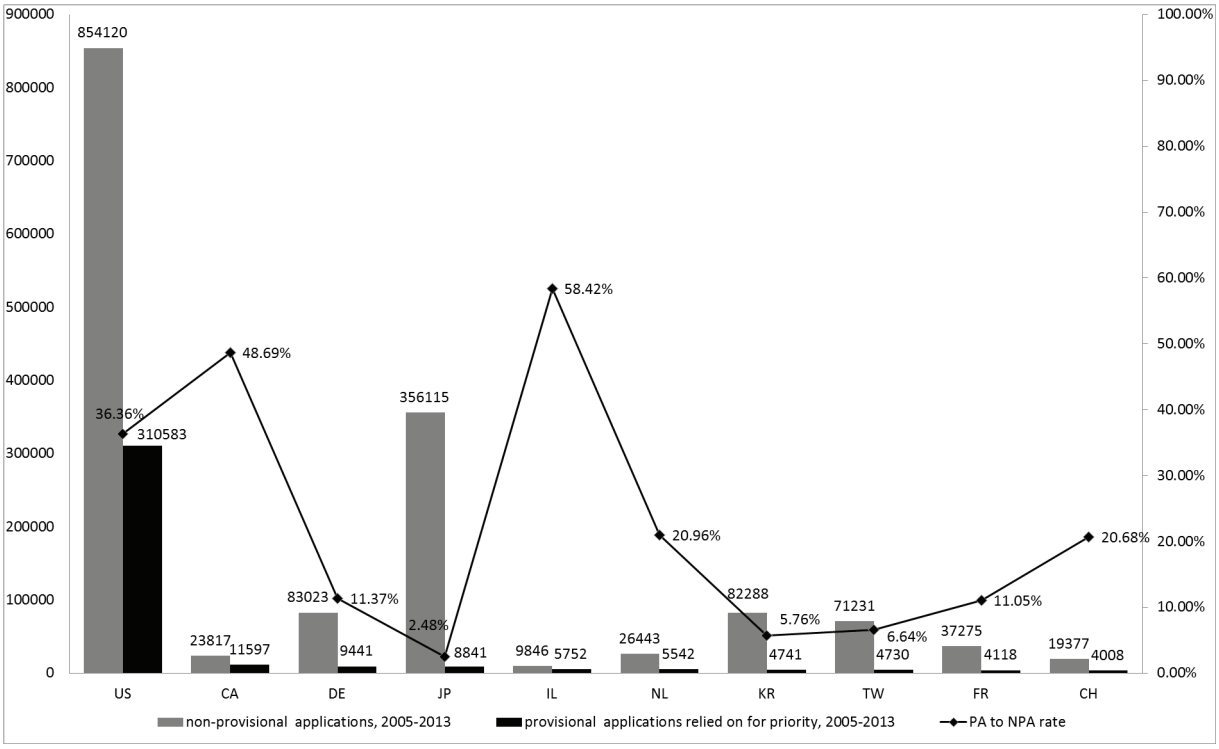


Figure 3. Top 10 countries of origins where applicants filed provisional applications with respect to corresponding non-provisional applications and the PA to NPA rate in the US in 2005-2013.

Furthermore, we checked the PA to NPA rate in order to find out the preference of filing provisional applications for applicants in different countries of origins. Figure 3 shows the top ten countries of origins, where applicants filed provisional applications with respect to corresponding non-provisional applications and the PA to NPA rate in the US in 2005-2013.

In Figure 3, the black bars represent the number of provisional applications filed by applicants from each country in the US in 2005-2013; the grey bars represent the number of non-provisional applications filed by applicants from each corresponding country in the US in 2005-2013; and the solid line represents the PA to NPA rate of each corresponding country in 2005-2013. Figure 3 shows that the PA to NPA rates of the US (36.36%), CA (48.69%) and IL (58.42%) were very much above the average percentage (about 17%). Contrarily, the PA to NPA rates of JP (2.48%), KR (5.76) and TW (6.64%) were far less than the average percentage. We reasoned that the English ability of applicants might be the cause for such a result. Comparing to applicants from JP, KR and TW, applicants from the US, CA and IL are either native English speakers or having good English abilities, so it is relatively easy for applicants in these countries to prepare a provisional application that is suitable for being

relied on for claiming priority by a non-provisional application. Moreover, some foreign laws limit the filing of patent applications abroad before a national patent application filing or authorization occurs. So the PA to NPA rate is expected to be low for applicants from those countries. For example, CN has this kind of law, and its PA to NPA rate was only 2.75%.

Provisional applications by different technological categories

In this paper, we used the six main technological categories (i.e. Chemical, Computers & Communications, Drugs & Medical, Electrical & Electronic, Mechanical, and Others) developed by The National Bureau of Economic Research (NBER) (Hall et al., 2001) to analyse provisional applications by technological categories.

Figure 4 shows the provisional applications relied on for priority filed each year from 2005 to 2013 divided by the NBER main technological categories. As shown in Figure 4, Computers and Communications and Drugs and Medical were the most popular main technological categories, in which applicants filed provisional applications and further converted them to non-provisional applications by claiming priority.

Sukhatme and Cramer (2014) suggested that an applicant who cares about the patent term will seize an opportunity to increase the term if it is offered to him/her. Applicants in industries in which the patent term is especially important would be more likely to file provisional applications than applicants in industries in which the term is less important. In the Drugs & Medical industry, the patent term is critical, i.e. applicants consider the patent term sensitivity, so the applicants tend to extend the statutory patent term up to one year by filing provisional applications first instead of non-provisional applications. In the Computers & Communications category, technologies change rapidly, i.e. applicants consider filing date sensitivity, so obtaining an early effective filing date is important to inventions in this category.

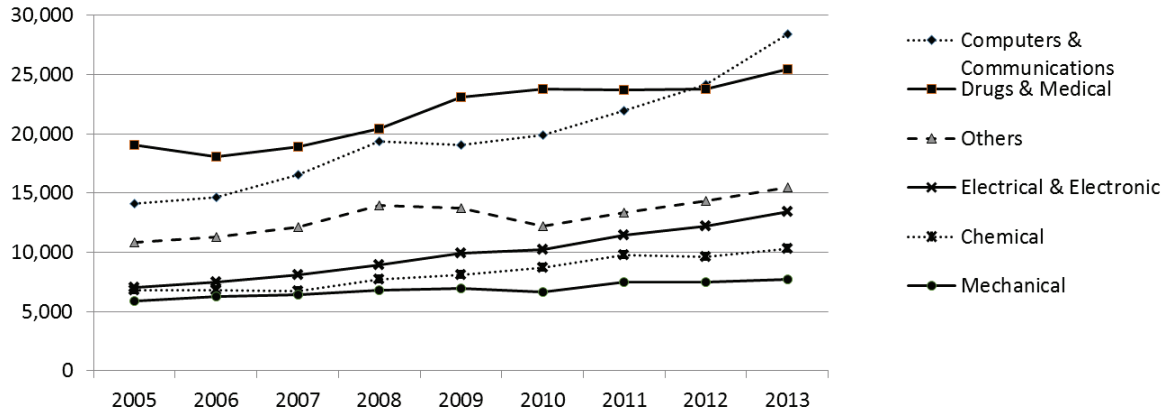


Figure 4. Provisional applications relied on for priority filed each year from 2005-2013, by NBER main technological categories.

Provisional applications by different assignees

Table 2 displays the top ten assignees filing provisional applications that were relied on for priority in the US in 2005-2013. Table 2 also shows the corresponding non-provisional applications by the top ten assignees, and their PA to NPA rates. It is clear that except for Samsung (5.68%) and Microsoft (9.27%), the PA to NPA rate of each of the other assignees was very much above the average percentage (about 17%). Take California University as an example, its PA to NPA rate was up to 81.28%. That is, in about every ten non-provisional

applications, over eight non-provisional applications claimed priority based upon early filing provisional applications.

Table 2. Top ten assignees filing provisional applications that were relied on for priority in the US in 2005-2013, the corresponding non-provisional applications, and the PA to NPA rates.

Assignee	*provisional applications relied on for priority*	*non-provisional applications*	*PA to NPA rate*
Qualcomm	6291	10018	62.80%
California University	3426	4215	81.28%
Broadcom	2876	4963	57.95%
Samsung Electro-Mechanics	2771	48814	5.68%
Koninklijke Philips Electronics N.V.	2519	12386	20.34%
Microsoft	2483	26799	9.27%
DuPont	2429	3286	73.92%
Texas Instruments	2353	5943	39.59%
LG Electronics	2318	9211	25.17%
Apple	1772	5124	34.58%

Table 3 shows main patent areas of each of the top ten assignees. For example, Qualcomm focused on the Computers & Communications field. So among all the 6291 provisional applications that relied on for priority, 5612 applications (about 89%) filed in the category of Computers & Communications. Broadcom, Samsung Electro-Mechanics, Microsoft, Texas Instruments, LG Electronics, and Apple also focused on the field of Computers & Communications.

Table 3. Provisional applications filed by the top ten assignees in the US in 2005-2013 by technological categories.

Assignee	*Chemical*	*Computers & Communications*	*Drugs & Medical*	*Electrical & Electronic*	*Mechanical*	*Others*
Qualcomm	0	5612	0	483	74	106
California University	514	269	1702	716	102	123
Broadcom	0	2264	0	441	0	145
Samsung Electro-Mechanics	0	2187	0	318	0	206
Koninklijke Philips Electronics N.V.	0	852	761	649	53	175
Microsoft	0	1880	0	107	0	474
DuPont	1007	0	509	394	95	374
Texas Instruments	0	1439	0	792	60	0
LG Electronics	0	2041	0	122	0	140
Apple	0	1169	0	429	38	107

It appears that applicants in the Computers and Communications field tend to file more provisional applications than those in other fields. We checked provisional applications that

were relied on for claiming priority filed by the top ten assignees in the Computers & Communications field each year between 2005 and 2013 and all provisional applications that were relied on for claiming priority filed by each of the top ten assignees each year between 2005 and 2013. The result was shown in Figure 5. For all the ten assignees, provisional applications filed in the Computers and Communications field were very close to all provisional applications. It indicates that, applicants in the Computers & Communications field only focused on one field.

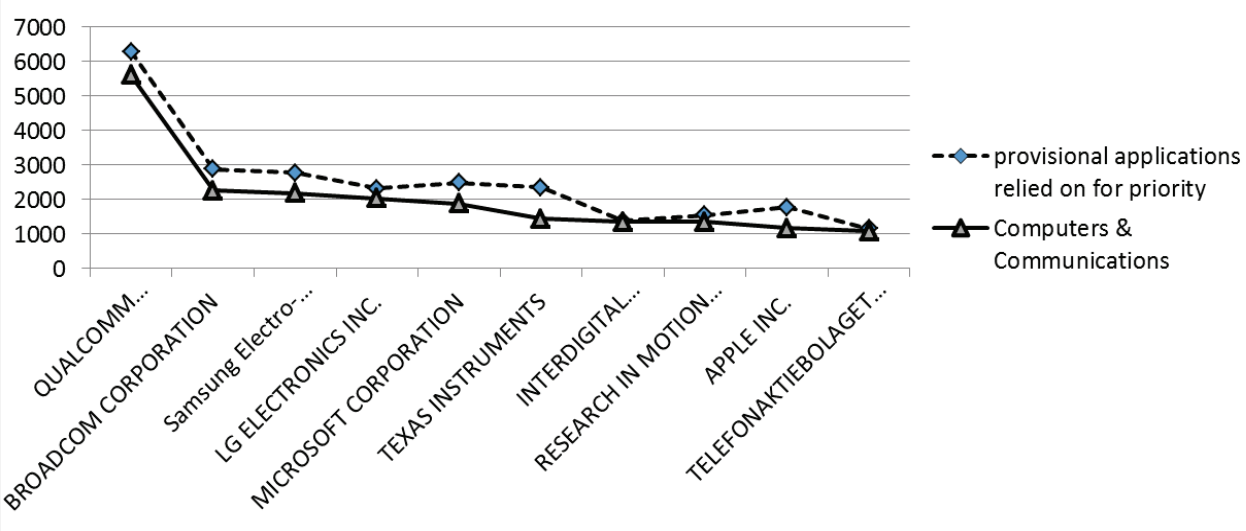


Figure 5. Provisional applications that were relied on for claiming priority filed by the top ten assignees in the Computers & Communications field each year between 2005 and 2013 and all provisional applications that were relied on for claiming priority filed by each of the top ten assignees each year between 2005 and 2013.

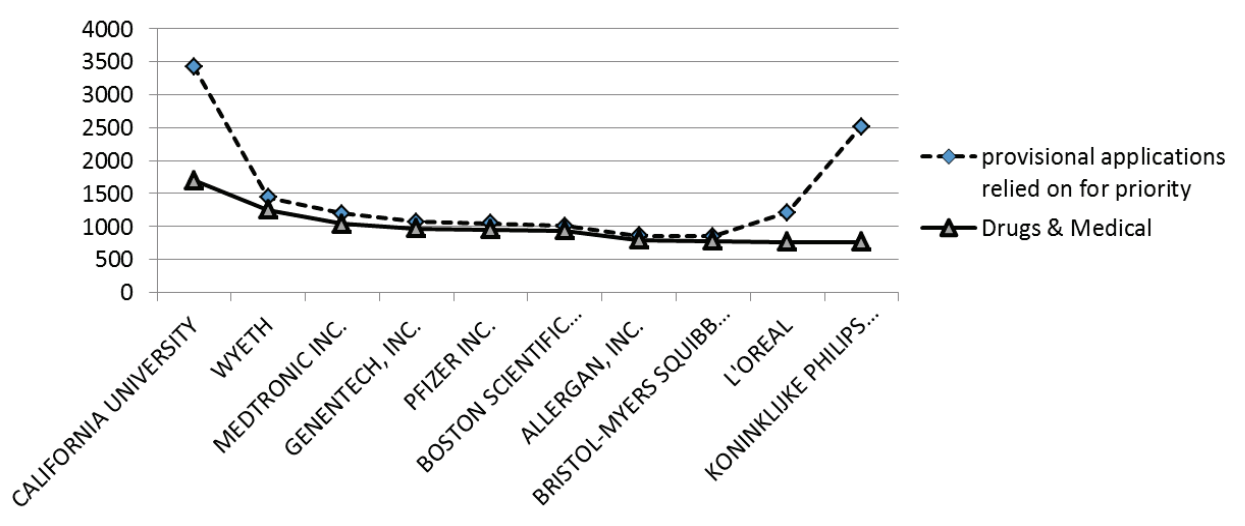


Figure 6. Provisional applications that were relied on for claiming priority filed by the top ten assignees in the Drugs & Medical field each year between 2005 and 2013 and all provisional applications that were relied on for claiming priority filed by each of the top ten assignees each year between 2005 and 2013.

Furthermore, we checked the provisional applications that were relied on for claiming priority filed by the top ten assignees in the Drugs and Medical field each year between 2005 and 2013 and all provisional applications that were relied on for claiming priority filed by each of the top ten assignees each year between 2005 and 2013. The result was shown in Figure 6.

Except for California University and Koninklijke Philips Electronics N.V., assignees filing provisional applications in Drugs & Medical also performed similarly to those in Computers & Communications, i.e. they had less diversity and only focused on one field.

Conclusion

It was found that provisional application filings continued to rise with an increase of non-provisional application filings between 2005 and 2013. The preference rate remained steady with a slight increase. The use rate of provisional applications was about 52% to 60% each year between 2005 and 2013. The PA to NPA rate can be used to evaluate the provisional application filings with respect to non-provisional application filings in different countries of origins, technological categories, and assignees. Filing date sensitivity, patent term sensitivity, and the necessity of promoting were regarded as factors weighing toward filing provisional applications. Cost sensitivity and English abilities were regarded as factors weighing against filing provisional applications.

For provisional applications by different countries of origins, applicants from Eastern Asian countries, including Japan, Korea, Taiwan and China, were less likely to file provisional applications in the US Contrarily, applicants form English speaking countries, including the US, Canada and Israel, were more likely to file provisional applications in the US. Therefore, applicants' English ability might be a major factor that influenced whether or not they would like to file provisional applications in the US.

For provisional applications by different technological categories, applicants in the fields of Computers and Communications and Drugs and Medical were more interested in filing provisional applications in the US.

For provisional applications by different assignees, most of the top ten assignees came from the Computers and Communications field.

References

Anderson, M.H., Cislo, D., Saavedra, J., & Cameron, K. (2014). Why International Inventors Might Want to Consider Filing Their First Patent Application at the United States Patent Office & the Convergence of Patent Harmonization and E-Commerce, *30 Santa Clara High Tech. L.J. 555*.

Crouch, D. (2008). *A First Look at Who Files Provisional Patent Applications.* Retrieved April 10, 2015 from: http://patentlyo.com/patent/2008/06/a-first-look-at.html

Crouch, D. (2012). *Provisional Patent Applications as a Flash in the Pan: Many are Filed and Many are Abandoned.* Retrieved December 31, 2014 from: http://patentlyo.com/patent/2012/11/provisional-patent-applications-as-a-flash-in-the-pan-many-are-filed-and-many-are-abandoned.html

Crouch, D. (2013). *Abandoning Provisional Applications*. Retrieved December 31, 2014 from: http://patentlyo.com/patent/2013/01/abandoning-provisional-applications.html

Crouch, D. (2014). *Claiming Priority to Provisional Applications.* Retrieved December 31, 2014 from: http://patentlyo.com/patent/2014/04/priority-provisional-applications.html

Hall, B. H., Jaffe, A. B., & Trajtenberg, M. (2001). The NBER patent citation data file: Lessons, insights and methodological tools (No. w8498). National Bureau of Economic Research.

Sukhatme, N.U & Cramer, J.N.L. (2014). *Who Cares About Patent Term? Cross-Industry Differences in Term Sensitivity.* Manuscript submitted for publication.

USPTO (2014). *Provisional Application for Patent.* Retrieved December 31, 2014 from: http://www.uspto.gov/ patents/ resources/types/provapp.jsp.

A Preliminary Study of Technological Evolution: From the Perspective of the USPC Reclassification

Hui-Yun Sung[1], Chun-Chieh Wang[2] and Mu-Hsuan Huang[3]

[1] *hsung@dragon.nchu.edu.tw*
Graduate Institute of Library and Information Science, National Chung Hsing University, Taichung (Taiwan)

[2] *chunchiehwang@ntu.edu.tw*
Department of Library and Information Science, National Taiwan University, Taipei (Taiwan)

[3] *Corresponding Author: mhhuang@ntu.edu.tw*
Department of Library and Information Science, National Taiwan University, Taipei (Taiwan)

Abstract
This study aimed to investigate technological evolution from the perspective of the USPC reclassification. The results showed that there existed significant differences among five types of patents based on the USPC reclassification: Patents reclassified to Class 001, Patents with Inter-field Mobilised Codes, Patents with Intra-field Mobilised Codes, Patents with Abolished Codes, and Patents with Original Codes. Patents reclassified to Class 001, mostly related to the topic of "Data processing", performed better than other patents in novelty, linkage to science, technological complexity and innovative scope. Patents with Inter-field Mobilised Codes, related to the topics of "Data processing: measuring, calibrating, or testing" and "Optical communications", involved broader technology topics but had a low speed of innovation. Patents with Intra-field Mobilised Codes, mostly in the Computers & Communications and Drugs & Medical fields, tended to have little novelty and a small innovative scope. Patents with Abolished Codes and patents with Original Codes performed similarly – their values of patent indicators were low. It is suggested that future research extend the patent sample to subclasses or reclassified secondary USPCs in order to understand the technological evolution within a field in greater detail.

Conference Topic
Patent Analysis

Introduction
For patented inventions, their technological novelty is indicated through their U.S. Patent Classification (USPC) assigned by the U.S. Patent Office. However, patent technology codes are an underutilized data resource for research on technological capabilities, technological novelty, technological complexity and technological change (Strumsky, Lobo & van der Leeuw, 2012). In order to fill the research gap, this study takes a first step towards using the USPC reclassification to trace technological evolution in the past two decades. This section introduces basic information regarding the USPC reclassification and sets out the research aim for investigation.

Reclassification of the U.S. Patent Classification (USPC)
The USPC is a system for organizing all U.S. patent documents and many other technical documents into relatively small collections based on common subject matter (USPTO, 2012b, I-1). A combination of a *class* (i.e. a major component) and a *subclass* (i.e. a minor component) is used to indicate every subject matter division in the USPC system. Based on the technology used, each patent is assigned specific USPC technology code(s) to reflect their technological topics. In order to distinguish from other patent classification schemes, this study only focuses on the USPC classification.

According to the USPTO (2012b, I-15), "[r]eclassification is the process of changing classifications assigned to documents classified in the USPC." There are different types of

modification of the USPC codes originally assigned to patents, including: creating, abolishing or modifying USPC class schedules. The USPC reclassification is seen necessary to reflect the evolving technological changes. For instance, Strumsky, Lobo and van der Leeuw (2012) used patent technology codes to study technological change.

Five types of patents based on the USPC reclassification

In order to keep pace with knowledge, modification/updates of classes and subclasses have been made to the Dewey Decimal Classification (DDC) system regularly. For instance, one of the new features in the DDC (Edition 23) was an update of "004–006 Computer science (and parallel provisions in 025.04 Information storage and retrieval systems and 621.39 Computer engineering) to reflect current technical trends" (Online Computer Library Center, 2013, p.3). Therefore, this study aims to investigate technological evolution from the perspective of the USPC reclassification.

As a result of the USPC reclassification, technology codes assigned to patents were created, modified and abolished. To this end, this study divided the utility patents into the following five types, according to the types of the modification of their original USPC:

- **Class 001:** If the record for a patent is incomplete and contains no *Primary Classification*[1], or if the USPTO is unable to assign specific technology codes to the patent, then the patent is reclassified to class 001, titled "CLASSIFICATION UNDETERMINED" (USPTO, 2012b).
- **Intra-field Mobilised Code:** A patent's newly assigned codes are derived from the same technological field as its original codes. Six technological fields are discussed in this paper, which are defined by Jaffe, Trajtenberg and Romer (2005).
- **Inter-field Mobilised Code:** A patent's newly assigned codes are derived from a different technological field from the original codes.
- **Abolished Code:** A patent's original technology codes are abolished and reclassified to new codes based on the Current USPC.
- **Original Code:** A patent's original technology codes remain the same as the newly assigned codes based on the Current USPC.

Based on the aforementioned five types of the utility patents, this study conducts a 20-year trend analysis and compares their variances using six patent indicators.

Methodology

Patent bibliometrics

In this study, patent data were collected solely from the United States Patent and Trademark Office (USPTO) database, which is generally accepted and is accessible to the researchers. While there exist different categories of patents (e.g. plant patents, design patents, reissues, and continuations), this study, based on the recommendations offered by Narin (2000), collected the number of regular U.S. utility patents to keep the focus of the database on the key category of patents, which contributes to corporate technological strengths. In order to observe the recent development of patents with the USPC reclassification, this study covered the past two decades. This study used the following six patent indicators to analyse the differences between different types of USPC reclassified patents.

- **Technology Cycle Time (TCT)** indicates the speed of innovation of a patent. Companies with a shorter cycle time than their competitors in a given technology area

[1] According to the USPTO (2012b), U.S. PGPub documents classified in the USPC are assigned one, and only one, principal mandatory classification, known as the *Primary Classification* (PR).

may be advancing more quickly from prior technology to current technology (Narin, 2000).

- **Non-Patent Reference (NPR)** indicates a patent's linkage to science. Narin (2000) proposed that the average rate of citations to scientific papers can be used to indicate the patent's science linkage. Other scholars (Gupta, 2006; Lo, 2010) also regarded the average rate of citations to NPRs as the patents' linkage to science. Therefore, this study used the number of NPRs to indicate the strength of linkage between the patent and science.
- **Patent Reference** indicates the novelty of a patent. A higher number of patent references generally indicate a reduction of invention novelty.
- **USPC Count** indicates the breadth of the technology topics of a patent. If a patent has broader technology topics, it tends to belong to a more highly applicable technological field.
- **Patent Term Extension** indicates the technological complexity of a patent. If the term of a patent is extended, it usually means that the patent involves a higher level of technological complexity and therefore requires more time for examination (Pantros IP, 2013).
- **Patent Claim** indicates the innovative scope of a patent. Patents containing a higher number of claims have been shown to have a wider innovative scope (Pantros IP, 2013).

Data collection

The empirical data analysed in this study were collected from the USPTO Granted Patent Database. The sample was restricted to the utility patents granted from 1994 to 2013. According to the classification system of Jaffe, Trajtenberg and Romer (2005), the U.S. patents were classified into six technological fields: Chemical, Computers and Communications (C&C), Drugs and Medical (D&M), Electrical and Electronics (E&E), Mechanical, and Others. The six fields were used to form the basis for an analysis of the patents with USPC reclassified inter-field or intra-field. USPC patents (with/without reclassification) were identified through the use of XML to compare Original USPC (i.e. USPC codes before reclassification) and Current USPC (i.e. USPC codes after reclassification). USPC reclassified patents in the recent 20 years were collected. In order to conduct a comparison analysis, the sample was randomly selected from the patents with Original USPC Codes that had the same patent count with Current USPC Codes each year.

Descriptive statistics

Descriptive statistics provide brief summaries about the sample and the observations made. Such summaries may be either quantitative (i.e. summary statistics) or visual (i.e. clear graphs). These summaries may either form the basis of the initial description of the data as part of a further statistical analysis, or they may be sufficient in and of themselves for a particular investigation. This study used the Line Chart to analyse the trends of patent counts for all types of the USPC reclassified patents granted each year. For the characteristic differences of each type of the USPC reclassified patents, this study used One-Way ANOVA to conduct significant difference tests on the patents' TCT, NPR, Patent Reference, USPC Count, Patent Term Extended, and Patent Claim.

In statistics, one-way analysis of variance (abbreviated one-way ANOVA) is a technique used to compare means of three or more samples (using the F distribution). The ANOVA tests the null hypothesis that samples in two or more groups are drawn from populations with the same mean values. To do this, two estimates are made of the population variance. If the group means are drawn from populations with the same mean values, the variance between the group means should be lower than the variance of the samples, following the central limit

theorem. A higher ratio therefore implies that the samples were drawn from populations with different mean values (Wikipedia, 2014).

Results

Trends of the USPC reclassified patents

There were 3,342,076 U.S. utility patents granted between 1994 and 2013. Among them, 102,204 patents belonged to the main class in Primary USPC reclassification, which accounted for 3.1% of the total utility patents. Calculations of those patents by their types showed that patents with Abolished Codes accounted for the majority (42.53%), which was followed by patents with USPC Intra-field Mobilised Codes. Patents with Class 001 or Inter-field Mobilised Codes accounted for appropriately 15% respectively. See Table 1.

Table 1. Counts of patents with/without USPC reclassification.

Patent with/without USPC Reclassification	Count
Main class in Primary USPC Reclassification	**102,204(100%)**
A. Class 001	15,862(15.52%)
B. Abolished Code	43,465(42.53%)
C. Inter-field Mobilised Code	15,740(15.40%)
D. Intra-field Mobilised Code	27,137(26.55%)
E. Random selection of patents with Original Code	**102,204**

Observed from the yearly distribution of the patent counts of various types of USPC reclassification, it was found that the number of USPC reclassified patents tended to be higher in the early stage, which indicated that the USPC was revised in accordance with the evolution of technologies. From the perspective of the Current USPC, some Original USPC appeared inappropriate in today's context and therefore the count of the USPC reclassified patents has increased. Furthermore, when the advance of newer technologies adopted the Original USPC that was similar to the version of October 2014, the number of USPC reclassified patents decreased in tandem.

The number of patents with Abolished Codes dramatically increased prior to 2000 but dramatically dropped after 2001, meaning that the elimination of main class did not occur after 2001. The number of patents with USPC Intra-field Mobilised Codes was above 1,000 before 2009 and started to decrease after 2010, which was considered relevant to "Technological development for stability". The numbers of patents with USPC Inter-field Mobilised Codes and with Class 001 tended to decrease in 2010, which was also considered relevant to "Technological development for stability".

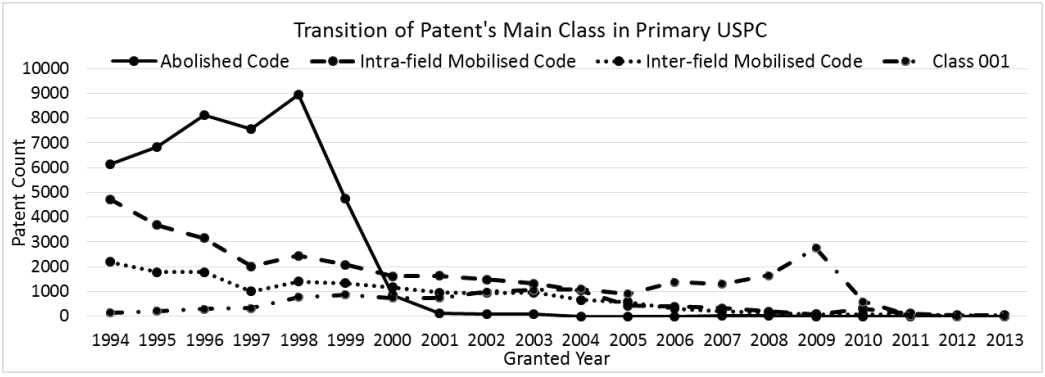


Figure 1. Transition of patents' main class in primary USPC.

Average citation rates were used to represent the quality of patents. This study calculated patents' average citation rates from 1994 to 2013, as shown in Figure 2. Due to the fact that the citation window of patents has become shorter each year, patents' average citation rates also decreased gradually. Figure 2 shows that the average citation rates of patents with Class 001 were the highest, which was followed by patents with USPC Inter-field/Intra-field Mobilised Codes. (They performed similarly in terms of their average cited rates recently.) The average citation rates of patents with Abolished Codes were higher than patents with Original Codes before 2002, but their average citation rates became the lowest among all types of patents.

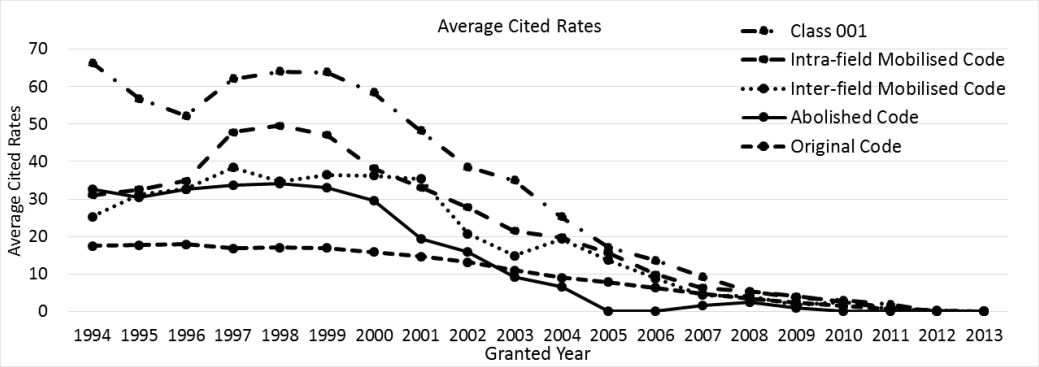


Figure 2. Average cited rates of USPC reclassified patents.

USPC reclassified patents among fields

Table 2. Patent counts in technological fields with USPC Reclassification.

	Patent Reclassified to Current Tech Field (%)						
Original Tech Field	**1.**	**2.**	**3.**	**4.**	**5.**	**6.**	**Sum**
1. Chemical	3,303	62	276	816	684	252	**5,393**
	(61.25)	(1.15)	(5.12)	(15.13)	(12.68)	(4.67)	**(100)**
2. Computer & Communication	135	11,649	16	1,201	81	1,913	**14,995**
	(0.90)	(77.69)	(0.11)	(8.01)	(0.54)	(12.76)	**(100)**
3. Drugs & Medical	958	13	6,260	44	23	96	**7,394**
	(12.96)	(0.18)	(84.66)	(0.60)	(0.31)	(1.30)	**(100)**
4. Electrical & Electronic	155	1,627	49	1,187	124	1,273	**4,415**
	(3.51)	(36.85)	(1.11)	(26.89)	(2.81)	(28.83)	**(100)**
5. Mechanical	979	3,037	74	172	2,773	237	**7,272**
	(13.46)	(41.76)	(1.02)	(2.37)	(38.13)	(3.26)	**(100)**
6. Others	756	94	111	159	323	1,965	**3,408**
	(22.18)	(2.76)	(3.26)	(4.67)	(9.48)	(57.66)	**(100)**
Sum	**6,286**	**16,482**	**6,786**	**3,579**	**4,008**	**5,736**	**42,877**
	(14.66)	**(38.44)**	**(15.83)**	**(8.35)**	**(9.35)**	**(13.38)**	**(100)**

Table 2 displays the U.S. utility patents granted from 1994 to 2013 with USPC reclassified inter/intra-field. It was found, through calculating the variances in the patent count in the original and current technological fields that patents in C&C were reclassified most among all the USPC reclassified patents. Among the patents in original technological fields in C&C, 77.69% belonged to the main class in the Primary USPC Intra-field Mobilised Code, with 12.76% reclassified to Others. Another variance occurred to D&M. 84.66% of the patents belonged to the main class in Primary USPC Intra-field Mobilised Code, with 12.76%

reclassified to Chemical. The last variance occurred to Mechanical. 38.13% of the patents belonged to the main class in Primary USPC Intra-field Mobilised Code, with 41.76% reclassified to C&C. 36.85% of patents in E&E were reclassified to C&C, 28.83% reclassified to Others, and only 26.89% reclassified intra-field.

Statistical differences among five patent groups

Six one-way between subjects ANOVAs were conducted to compare the effect of patents with different USPC reclassification types on patent performance in TCT, NPR, Patent Reference, USPC Count, Patent Term Extended, and Patent Claim. There were all significant differences of indicators on patent performance at the $p<.001$ level for the five types of patents with/without USPC reclassification. Post hoc comparisons using the Dunnett T3 test (Dunnett, 1980) showed significant differences in the mean scores of the six indicators for the patents in different types of the USPC reclassification.

- **TCT Performance:** When the value of TCT is lower, it means a patent involves more fast-moving technologies and a patent tends to cite recently issued patents. Results derived from statistical tests showed: B. Abolished Code (5.7 year) < C. Inter-field Mobilised Code (6.3 year) < E. Original Code. (7.8 year). Short TCT of the patents with Abolished Codes indicated that patents of this kind involved the most fast-moving technologies and the speed of their technological innovation was clearly faster than patents with Inter-field Mobilised Codes. On the contrary, patents with Original Codes tended to be slower in term of their speed of the technological innovation.
- **NPR**: When the number of NPR is higher, it means the linkage of technology to science is stronger. Results derived from statistical tests showed: A. Class 001 (10.4), C. Inter-field Mobilised Code (11.7) & D. Intra-field Mobilised Code (10.7) > E. Original Code (7.9) & B. Abolished Code (5.5). When calculating Science Linkage, the more NPRs were, the stronger the linkage of technology to science was. Therefore, patents reclassified to Class 001, patents with Inter-field Mobilised Codes and Intra-field Mobilised Codes had stronger linkages to science, compared to patents with Original Codes and Abolished Codes.
- **Patent Reference:** When the number of Patent References is low, it indicates the novelty of technology is high. Results derived from statistical tests showed: B. Abolished Code (11.6) < E. patent with Original Code (14.2) < A. Class 001 (19.3) < C. Inter-field Mobilised Code (15.0). It can be inferred that the technological novelty of patents with Abolished Codes was much higher than that of patents with Original Codes. Clearly, the technological novelty of patents with Class 001 or with Inter-field Mobilised Codes tended to be low.
- **USPC Count:** Patents with more USPC counts indicate they involve broader technologies. Results derived from statistical tests showed: C. Inter-field Mobilised Code (5.2) > E. Original Code (4.4) > B. Abolished Code (3.9). The technology breadth of patents with Inter-field Mobilised Codes was the largest. The technology breadth of patents with Abolished Codes was smaller than that of patents with Original Codes.
- **Patent Term Extended:** When the term extension lasts longer, it indicates that a patent involves more complicated technologies. Results derived from statistical tests showed: A. Class 001 (416) > C. Inter-field Mobilised Code (341), D. Intra-field Mobilised Code (307) > E. patent with Original Code (300) > B. Abolished Code (168). It can be inferred that patents with Class 001 involved a higher level of technological complexity than patents with Inter/Intra-field Mobilised Codes. However, the term extension of patents with Abolished Codes was the shortest, indicating that they involved the lowest level of technological complexity.

- **Patent Claim:** When the value of patent claims is higher, it indicates that a patent's innovation scope is wider. Results derived from statistical tests showed: A. Class 001 (22.2) > C. Inter-field Mobilised Code (17.6) > B. Abolished Code (16.5), E. patent with Original Code (15.1). It can be inferred that the innovation scope of the patents with Class 001 or patents with Inter-field Mobilised Codes was obviously wider than that of patents with Abolished Codes and patents with Original Codes.

Technological evolution from the USPC reclassification perspective

This study divided patents granted in the last two decades into two groups, i.e. 1994-2003 and 2004-2013. Observations were made from the evolution of USPC codes as a result of the USPC reclassification. Table 3 shows the USPC with top three most patent counts in the two periods respectively. If a patent was reclassified to Class 001, it meant that there was no specific technology code suitable for the patent. To some extent, it indicated that the patent belonged to emerging technologies or original USPC codes assigned were not appropriate for the patent, which required a new code. Table 3 shows in both periods, the majority of patents reclassified to Class 001 came from Class 707 in the C&C field. This phenomenon reflected the technological uncertainty of patents originally assigned to Class 707, the majority of which were therefore reclassified to Class 001. In the first period, there were 19.7% of patents originally assigned to Class 395 and then reclassified to Class 001. However, due to the abolition of Class 395, their technological description remained unknown.

Table 3. Patents with USPC reclassified in the Class 001 and the Abolished Code groups.

USPC	1994-2003	2004-2013	USPC Description
Original class reclassified to 001 (Class 001)			
707	4,884 (79.6%)	9,684 (99.5%)	Data processing: database and file management or data structures
395	1,206 (19.7%)	0 (0.0%)	(Abolished)
364	19 (0.3%)	0 (0.0%)	(Abolished)
705	0 (0.0%)	18 (0.2%)	Data processing: financial, business practice, management, or cost/price determination
714	0 (0.0%)	7 (0.1%)	Error detection/correction and fault detection/recovery
Current class of original abolished (Abolished Code)			
438	4,895 (11.3%)	1 (4.3%)	Semiconductor device manufacturing: process
714	4,179 (9.6%)	0 (0.0%)	Error detection/correction and fault detection/recovery
710	3,448 (7.9%)	0 (0.0%)	Electrical computers and digital data processing systems: input/output
703	1,314 (3.0%)	2 (8.7%)	Data processing: structural design, modeling, simulation, and emulation
477	2 (0.0%)	2 (8.7%)	Interrelated power delivery controls, including engine control

For patents with Abolished Codes, it meant that their original codes did not align with the technological evolution any more, and thus the codes were abolished and the patents were reclassified to new codes. As shown in Table 3, the majority of patents with Abolished Codes

occurred in the first period, with only 23 patents of this kind in the second period. In the first period, the majority of patents whose original USPC codes were abolished were reclassified to Classes 438 (11.3%), 714 (9.6%), 710 (7.9%), and 703 (3.0%). Patents reclassified to Class 438 were about semiconductor device manufacturing in the E&E field, and those reclassified to Classes 714, 710 and 703 focused on technologies in the C&C field. Based on the patents reclassified to Class 001 and with Abolished Codes, it was found that the USPC reclassification tended to occur in the C&C and E&E fields in the first period and in the C&C field in the second period.

According to Table 2, patents with Intra-field Mobilised Codes mainly occurred in the C&C (77.69%) and D&M (84.66%) fields. Therefore, Table 4 focuses on the top three Intra-field Mobilised Codes, and Figures 3 and 4 present the flow of the patents between USPCs in the two fields, where the flow occurred more than ten patents. In the C&C field, the USPC reclassification in both periods mainly occurred from Class 345 to Class 715 (28.8% and 26.6%), which was about "Operator interface processing" and from Class 369 to Class 720 (11.8% and 5.1%), which was about "Information storage or retrieval". Additionally, in the first period, there remained 10.2% of patents reclassified from Class 707 to Class 715, which was also about "Operator interface processing". In the second period, there remained 6.2% of patents reclassified from Class 707 to Class 709, which was about "Multicomputer data transferring". In the D&M field, the USPC reclassification occurred from Class 128 to Class 600 (68.0%) which was about "Surgery" in the first period, and from Class 514 to Class 424 (76.4%) which was about "Drug, bio-affecting and body treating compositions" in the second period. The code mobilisation within the same field occurred due to the extension of the original USPC.

Table 4. USPC reclassification: the Intra-field Mobilised Code group.

Main Class of USPC		Count	
Original	**Current**	**1994-2003**	**2004-2013**
Intra-field Mobilised Code in C&C			
345	**715**	2,793(28.8%)	516(26.6%)
369	**720**	1,144(11.8%)	98(5.1%)
707	**715**	991(10.2%)	96(5.0%)
707	**709**	68(0.7%)	120(6.2%)
Intra-field Mobilised Code in D&M			
128	**600**	3,909(68.0%)	4(0.8%)
514	**424**	626(10.9%)	389(76.4)
606	**623**	227(3.9%)	18(3.5%)
514	**435**	17(0.3%)	26(5.1%)
435	**424**	49(0.9%)	21(4.1%)

345: Computer graphics processing and selective visual display systems; **369:** Dynamic information storage or retrieval; **707:** Data processing: database and file management or data structures; **709:** Electrical computers and digital processing systems: multicomputer data transferring; **715:** Data processing: presentation processing of document, operator interface processing, and screen saver display processing; **720:** Dynamic optical information storage or retrieval

128: Surgery; **424:** Drug, bio-affecting and body treating compositions; **435:** Chemistry: molecular biology and microbiology; **514:** Drug, bio-affecting and body treating compositions (an integral part of Class 424); **600:** Surgery (an integral part of Class 128); **606:** Surgery (an integral part of Class 128); **623:** Prosthesis (i.e., artificial body members), parts thereof, or aids and accessories therefor

Observed from the patents with Intra-field Mobilised Codes, it showed that in the C&C field those patents were related to "Operator interface processing" in both periods. In the D&M field those patents were related to "Surgery" in the first period and "Drug, bio-affecting and body treating compositions" in the second period. Observed from the patents with Inter-field Mobilised Codes, it showed that the USPC codes were mainly mobilised from the E&E and Mechanical fields to the C&C field, as seen in Table 2. Statistics on the top three USPC mobilisation were detailed in Table 5, and Figures 5 and 6 present the flow of the patents between USPCs among the three fields, where the flow occurred more than ten patents. In the first period, the USPC reclassification mainly occurred from the E&E field to the C&C field, for example from Class 348 to Class 375 (64.6%) about "Pulse or digital communications", and from Class 346 to Class 374 (20.6%) about "Thermal measuring and testing". However, in the second period, inter-field code mobilisation was not obvious. It can be seen that the topics of technological evolution were different in the two periods.

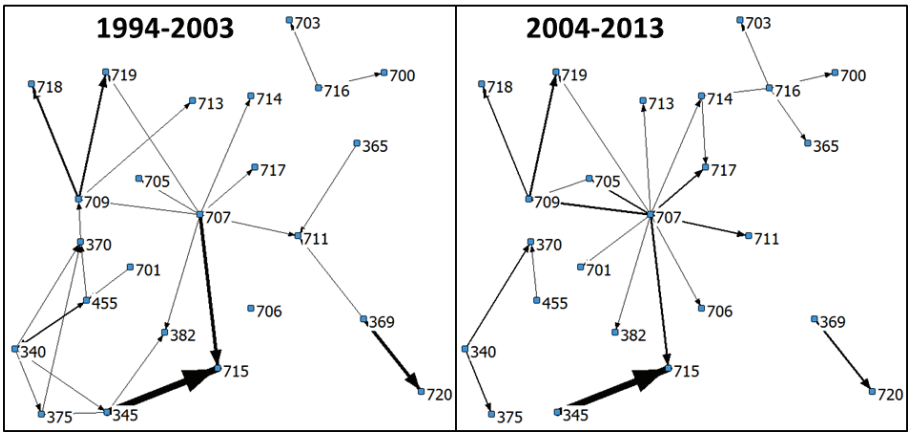


Figure 3. The flow of patents between USPCs in the C&C field.

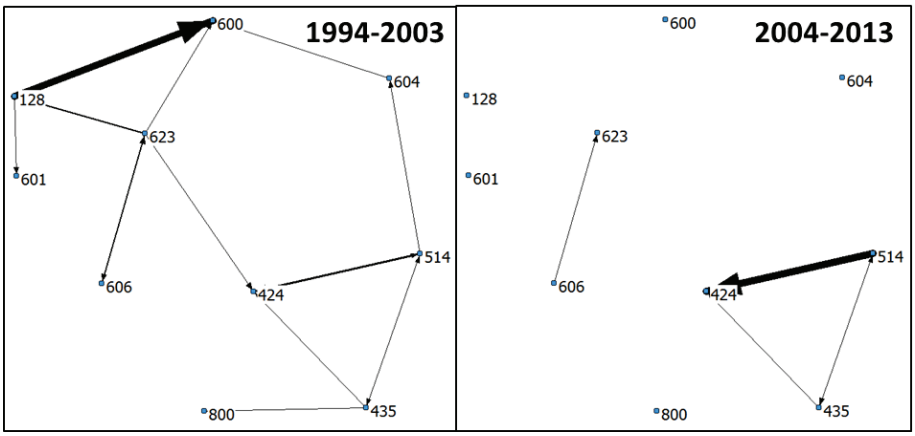


Figure 4. The flow of patents between USPCs in the D&M field.

Looking at patents with Inter-field Mobilised Codes from the Mechanical field to the C&C field, the flow of the mobilisation tended to occur from Class 359 to Class 398 (94.8% and 37.8%) about "Optical communications" in both periods.

Observed from the patents with Inter-field Mobilised Codes, it showed that patents with the USPC reclassification from the E&E field to the C&C field focused on the technology topics of "Pulse or digital communications" and "Thermal measuring and testing" in the first period, but focused on "Data processing: measuring, calibrating, or testing" in the second period. As

for patents with USPC reclassification from the Mechanical field to the C&C field, they tended to be related to "Optical communications" in both periods.

Table 5. USPC reclassification: the Inter-field Mobilised Code group.

Main Class of USPC		Count	
Original	**Current**	**1994-2003**	**2004-2013**
Inter-field Mobilised Code from E&E to C&C			
348	**375**	989(64.6%)	2(2.1%)
346	**374**	316(20.6%)	0(0.0%)
257	**365**	21(1.4%)	3(3.2%)
257: Active solid-state devices (e.g., transistors, solid-state diodes); **346:** Recorders; **348:** Television; **365:** Static information storage and retrieval; **374:** Thermal measuring and testing; **375:** Pulse or digital communications			
Inter-field Mobilised Code from Mechanical to C&C			
359	**398**	2,837(94.8%)	17(37.8%)
235	**705**	22(0.7%)	0(0.0%)
359	**369**	15(0.5%)	0(0.0%)
235: Registers; **359:** Optical: systems and elements; **369:** Dynamic information storage or retrieval; **398:** Optical communications; **705:** Data processing: financial, business practice, management, or cost/price determination			

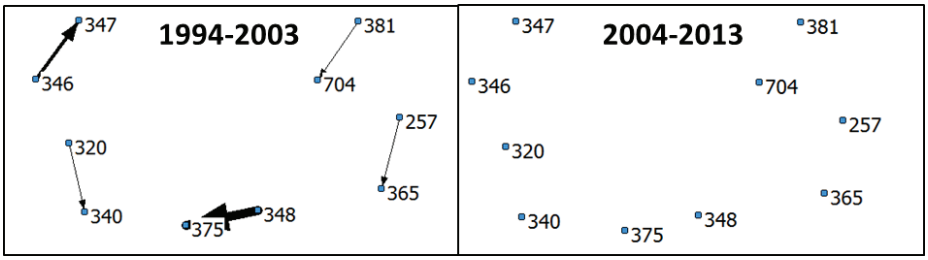


Figure 5. The flow of patents between USPCs from the E&E to the C&C field.

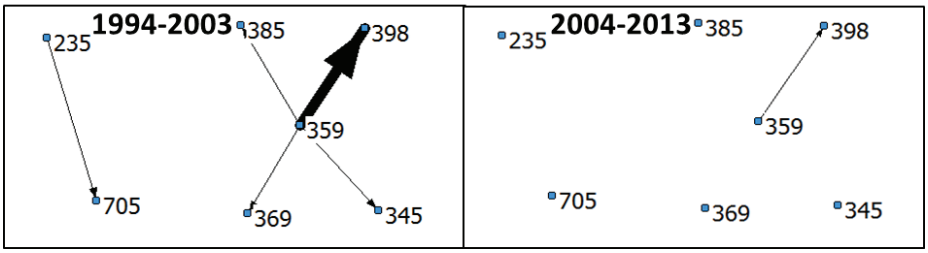


Figure 6. The flow of patents between USPCs from the Mechanical field to the C&C field.

Conclusion and Discussion

The majority of USPC reclassified patents occurring prior to 2000 and in the Computer & Communications field

With the advance of new technologies, the USPC system is updated quarterly in March, June, September and December (USPTO, 2012a). Newly granted patents were assigned with technology codes derived from the latest version of the USPC. Accordingly, their original USPC technology codes were less likely to be reclassified. This study found that the number of patents with main class in primary USPC reclassification hit the highest prior to 2000 and began to decrease every year after 2001. Patents with Abolished Codes accounted for 42.53%

and the majority of the patents were granted prior to 2000. Next were patents with Intra-field Mobilised Codes, which accounted for 26.55%. For the average citation rates every year, patents reclassified to Class 001 were ranked as top, and patents with Original Codes were ranked as bottom. Due to the USPC reclassification, patents with Intra-field Mobilised Codes occurred most frequently in the C&C field, and patents with Inter-field Mobilised Codes occurred most frequently from the Mechanical field to the C&C field.

USPC reclassified patents showing significant differences in patent indicators

Six one-way between subjects ANOVAs were conducted to compare the effects of patents in different groups by the USPC reclassification, according to their patent performance in TCT, NPR, Patent Reference, USPC Count, Patent Term Extended, and Patent Claim. Different results were obtained for the different types of patents, as below.

- **Patents reclassified to Class 001:** They got higher values of NPR, Patent Reference, Patent Term Extended and Claims Count, indicating that they performed better than other patents (whether they were reclassified or not) in novelty, linkage to science, technological complexity and innovative scope. Therefore, USPTO needs to re-examine appropriate USPC technology codes for them or assign appropriate codes to them when the new codes are created.
- **Patents with Inter-field Mobilised Code**: Compared to patents reclassified to Class 001, they got more USPC counts and longer TCT, indicating that they involved broader technology topics and therefore their codes assigned were mobilised inter-field. Their longer TCT meant that their technology had a low speed of innovation.
- **Patents with Intra-field Mobilised Code**: They tended to have low novelty and a small innovative scope; therefore, their codes assigned were mobilised intra-field.
- **Patents with Abolished Code**: They were mainly granted prior to 2000. Patens of this type and patents with Original Code performed similarly – their values of patent indicators were low.

Technological evolution from the perspective of the USPC reclassification

This study investigated different groups of patents based on the USPC reclassification. Statistical analysis was conducted on the technology codes and comparisons were made between two ten-year periods. Based on the results derived, different types of technological evolution were found.

- **Emerging technologies in Class 001:** In both periods, a large portion of the emerging technologies were about "Data processing: database and file management or data structures" in the C&C field. This reflects the uncertainty of the development of the emerging technology, and thus patents originally assigned to Class 707 needed to be continually redefined and reassigned with specific technology codes.
- **Technological transition in Inter-field Mobilised Code:** Technologies from the E&E and Mechanical fields tended to be transferred and applied to the C&C field. Technologies about "Television" in E&E was transferred and applied to "Pulse or digital communications" in the C&C field. Technologies about "Recorders" in E&E were also transferred and applied to "Thermal measuring and testing" in the C&C field. In the Mechanical field, technologies related to "Optical: systems and elements" were transferred and applied to "Optical communications" in the C&C field in both periods.
- **Technological cohesion or spread in Intra-field Mobilised Code:** Technologies in this group tended to focus on the C&C and D&M fields. In the C&C field, technologies related to "Computer graphics processing and selective visual display systems" and "Data processing: database and file management or data structures" were combined together and applied to "Data processing: presentation processing of document, operator

interface processing, and screen saver display processing". Figure 3 shows not only technological cohesion but also technological spread. For example, technologies about "Data processing: database and file management or data structures (Class 707)" were spread to other technologies in different fields. Patents with original USPC 707 were reclassified to eight different codes in the first period, and then spread to other ten codes in the second period.

- **Technological substitution in Abolished Code:** Technologies in this group tended to occur in the first period. This indicates that the USPC scheme in the second period has been adapted to the recent technological development. In the first period, technologies of this kind mainly occurred to those related to "Semiconductor device manufacturing", which were reclassified to Class 438 with their original USPC 437 being abolished. Technologies related to "Error detection/correction and fault detection/recovery" which were reclassified to Class 714 with their original USPC 371 and 395 being abolished. This indicates that the mature technologies have caused the biggest impact on the USPC scheme.

It is suggested that future research extend the sample to patents with reclassified USPC subclasses or patents with reclassified secondary USPCs in order to observe recent intra-field technological changes in great detail. The Radical (Leaps) Innovation of technologies is only applied to the minority, but the majority of patents are embedded with Incremental Innovation. Incremental Innovation tends to occur inside fields. Through extending the patent sample to subclasses or secondary of USPC, it helps understand more technological evolution within a field. Besides, understanding the establishment, abolishment and movement of technology codes recorded in the Classification Orders Archival Report (USPTO, 2013) helps understand the trajectories of technological evolution more detail. Although this study focused on the reclassification of USPC schemes, it is argued that the same research model could be applied to trace the changes in the class schemes in International Patent Classification (IPC) or Cooperative Patent Classification (CPC) and changes in classification codes in their counterpart patents.

References

Dunnett, C. W. (1980). Pairwise Multiple Comparisons in the Homogeneous Variance, Unequal Sample Size Case. *Journal of the American Statistical Association*, 75, 789-795.

Gupta, V. K. (2006). References to literature in patent documents: A case study of CSIR in India. *Scientometrics*, *68*(1), 29-40.

Lo, S. S. (2010). Scientific linkage of science research and technology development: A case of genetic engineering research. *Scientometrics*, *82*(1), 109-120.

Narin, F. (2000). Tech-line® background paper. In Tidd J. (Ed.), *From knowledge management to strategic competence* (pp. 155–195). London: Imperial College Press.

Online Computer Library Center. (2013). New Features in DDC Edition 23. Retrieved April 10, 2015 from: https://www.oclc.org/content/dam/oclc/dewey/versions/print/new_features.pdf

Pantros IP. (2013). Patent Factor Reports. Retrieved January 7, 2015 from: http://admin.patentcafe.com/reports/pantrosip_reports/patentfactor_terms.pdf

Strumsky, D., Lobo, J., & van der Leeuw, S. (2012). Using patent technology codes to study technological change. *Economics of Innovation and New Technology*, *21*(3), 267-286.

USPTO. (2012a). Classification Orders Index (COI). Retrieved January 10, 2015 from: http://www.uspto.gov/patents/resources/classification/orders/coi.jsp.

USPTO. (2012b). Overview of the U.S. Patent Classification System (USPC). Retrieved January 7, 2015 from: http://beta.uspto.gov/sites/default/files/patents/resources/classification/overview.pdf.

USPTO. (2013). Classification Order Archival Report. Retrieved January 7, 2015 from: http://www.uspto.gov/patents/resources/classification/archiverpt.pdf.

Wikipedia. (2014). One-way analysis of variance. Retrieved January 7, 2015 from: http://en.wikipedia.org/wiki/One-way_analysis_of_variance.

Cognitive Distances in Prior Art Search by the Triadic Patent Offices: Empirical Evidence from International Search Reports

Tetsuo Wada

tetsuo.wada@gakushuin.ac.jp
Gakushuin University, Faculty of Economics, Mejiro, Toshima-ku, Tokyo 171-8588 (Japan)

Abstract
Despite large numbers of empirical studies are conducted on examiner patent citations, few have scrutinized the cognitive limitations of officials at patent offices in searching for prior art to add citations during patent prosecution. This research takes advantage of the longitudinal gap between International Search Reports (ISRs) required by the Patent Cooperation Treaty (PCT) and subsequent examination procedure in national phase. It inspects whether several kinds of distances actually affect the probability that a piece of prior art is caught at the time of ISRs, which is much earlier than national phase examinations. Based on triadic PCT applications for all of the triadic patent offices (EPO, USPTO, and JPO) between 2002 and 2005 and their citations made by the triadic offices, evidence shows that geographical and organizational distances negatively affect the probability of prior patents being caught in ISRs, while lag of prior art positively affects the probability. Also, technological complexity of an application negatively affects the probability, whereas the size of forward citations of prior art affects positively.

Conference Topic
Patent Analysis (foundation of examiner patent citations, in particular)

Introduction
Patent citations have been widely utilized for empirical studies of patent systems, particularly for such issues as economic value and knowledge flows. Several empirical studies have examined whether examiner citations are different from inventor citations. One of the studies on the subject was conducted by Alacer and Gittleman (2006), who showed the similarity between examiner citations and inventor citations with respect to geographical distance in particular. While previous studies have compared examiner citations and inventor citations in other aspects such as the relationship with renewal rates, there have not been enough analyses concerning how patent offices are influenced by several kinds of "distances" that can limit cognitive boundary during prior art search. This study focuses on ISRs as a basis for measuring the search obstacles of the triadic patent offices, and tests how officials are bounded by "distances," including similar kinds of cognitive obstacles against prior art search, without relying on comparison with inventor citations. In conducting the analyses, we consider applicants' self-selection, since applicants from the U.S. and Japan can choose the European Patent Office as their search agency, where the EPO has reputation for its complete search (applicants who seek stringent search may choose the EPO ex ante).

The methodology: PCT and ISR as the basis of empirical measurement
This project proposes and implements a method of measuring the search obstacles, namely binding conditions on search capability, of the triadic patent offices by focusing on ISRs issued by different ISAs, specifically the patent offices in Europe, the U.S. and Japan, according to the PCT. In particular, binary choice models are employed for each of cited patents (which are added in the national phase in all of three jurisdictions) about whether or not they were already caught at the earlier time of ISR issued by the triadic offices. We limit our samples to those PCT applications made to and examined at all of the three offices. There are advantages to employ this methodology.

First, ISRs are issued under the common search criterion imposed by the WIPO under the PCT system. Under the PCT, "an applicant must file an application with a receiving office and choose an international searching authority to provide an international search report and a written opinion on the potential patentability of the invention." "The applicant generally has at least 30 months from the filing (priority) date to decide whether to enter the national phase in the countries or regions in which protection is sought" (WIPO, 2014). The guideline at the WIPO applies to every ISA when issuing ISRs, whereas applicants in some countries are allowed to choose ISAs. The same criterion for prior art search is applied over different patent offices, while national phase examinations do not have such standardized rules.

Second, the lag mentioned above between ISRs and national phase examinations allows a "level" testing ground for search completeness. While ISRs are issued at an early stage, more searches are conducted in national offices later. Since knowledge is geographically localized (Jaffe et al. 1993; 1999), and knowledge diffusion takes time, additional time between ISRs and national phase search facilitates more complete search in the later stage. We limit our samples to those PCT applications that are examined at all of the three triadic offices, meaning that localized knowledge in any of these areas at the time of ISRs is more likely to be caught by the offices at the national phase in a less localized way. See Figure 1 below for the lag and collective searches made at later stages in national phase.

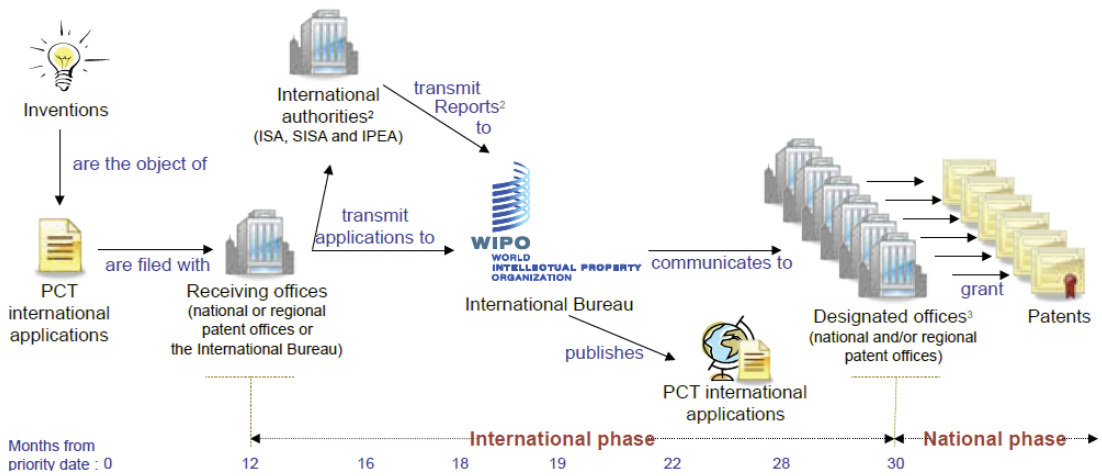


Figure 1. PCT procedure (replicated from WIPO, 2014, p.13).

Following the logic above, we retrospectively define the probability of every cited patent depicted in national phase, identified at the INPADOC family level, to have been already caught in the ISR of the originating PCT application. Taking this probability (a binary variable *found_in_ISR*, empirically) as the dependent variable, we implement PROBIT analyses at INPADOC family level with explanatory variables representing the various "distances" between citing and cited patents, including technological complexity of originating applications, and other related indicators.

Applicants' (inventors') citations are excluded from the analysis, since the objective is to evaluate the determinant of search completeness by the ISAs. However, self-selection of the U.S. and Japanese applicants to choose the EPO as their ISA is considered in the analyses, since the EPO has high reputation of examination standard and therefore applications with higher quality from the U.S. and Japan may choose the EPO as the ISA.

Although actual ISR search is sometimes outsourced to non-PTO agencies, we consider ISRs as a basis of evaluating PTOs, since they are issued under the name of the patent offices, not private search agencies. Only citations made by the triadic offices are considered in the current analyses. Since PATSAT, our primary data source, records non-patent literature in

non-standardized formats, we could not consolidate the same non-patent literature across different records. For this reason, we employ patent citations only at this time.

Hypotheses

Since ISR searchers (examiners/searchers for patent offices) are affected by cognitive obstacles from various "distances," we hypothesize that a prior patent (that was found in ISR or national phase) is more likely to be found in ISR when "distances" are less problematic, i.e.,
H1) a relevant prior patent is closer in geography (physical distance),
H2) prior patent is older (knowledge diffusion time),
H3) prior patent is from the same applicants (organizational distance),
H4) prior patent has more number of forward citations (knowledge diffusion probability), and
H5) application for which an ISR is issued has less scope, less number of claims, less number of inventors, and less number of international family (complexity against diffusion).
In addition, we consider if applicants' self-selection of ISAs affects the outcome variable.

Data source

The empirical domain of analysis is the triadic patent applications through PCT, with their earliest priority date within its international family between 2002 and 2005. Triadic PCT patent applications are defined here as INPADOC families that contain all of EPO, USPTO and JPO applications recorded on EPO's PATSTAT database, with only one "WO (PCT)" application in a family, meaning that a single PCT application initiates international phase for all applications in a family. The number of international families for the analysis is 97,828. Although international applications to and from China and Korea has increased dramatically in the last ten years, the triadic patent offices of the EPO, the USPTO and the JPO represented the vast majority before 2005, which is our observation period.
EPO PATSTAT (2013 OCT version) is used, and INPADOC family is the unit of analysis. Citation data also comes from PATSTAT (2013 OCT), although JPO citation data is augmented by Seiri-Hyojunka data (JPO's standardized patent prosecution data). US citations are not complete as well on PATSTAT, since citations for rejected applications are not registered on PATSTAT. The lack of the US citations for rejected applications may affect the result of the analysis, but this has not been verified yet. Applicant identifiers are consolidated by the EEE-PPAT database developed by ECOOM (Du Plessis et al., 2009; Magerman et al., 2009; Peeter et al., 2009).

Variables

We employ several categories of explanatory variables, representing each of hypotheses above, in PROBIT analyses taking the probability of a cited patent being caught in the previous ISR as the binary dependent variable ("*found_in_ISR")*. The unit of analysis is a pair of citing and cited international families, both consolidated at INPADOC family level.
For H1, three variables of *euro_cited* (cited family has its 1st priority, i.e., the earliest date, in EPC countries within a family, derived from tls201 and tls219 tables of PATSTAT), *us_cited* (cited family has its 1st priority in the U.S.), and *jp_cited* (cited family has its 1st priority in Japan) are defined. When a cited family has its origin in the same region where ISR is issued, the ISA of the region is expected to have geographical advantage over the relevant technology. Expected sign is positive for each region, e.g., positive *jp_cited* coefficients for applications originating from Japan.
For H2, citation lag between the 1st priority of a citing family and that of a cited family is defined as *fam_cite_lag* (derived from tls201 and tls219 tables of PATSTAT). The longer the lag is, the easier the prior art will be to be found at the time of ISR.

For H3, *self* is defined as a binary variable, taking the value of one if one of patents in a cited family and one of patents in a citing family belongs to the same applicant, based on PATSTAT (tls207) combined with EEE-PPAT, using "L2" id. Patent office will find it easier to locate prior relevant art within the same applicant.
For H4, *fwd_cite_of_the_cited* is defined and obtained from PATSTAT (tls217) as the number of forward examiner citations, counted at publication level (but consolidated at family level), and made out to the cited patent family.
For H5, we first use scope indicators. *IPC4_count* is the total net count of IPC subclasses (4-digit IPC, derived from tls209) assigned in a citing INPADOC family. Since patent classification of an application may change during prosecution process both in international phase and in national phase, we include all IPC subclasses to capture the breadth of a family. The number of claims of a patent is correlated with the complexity of the technological content. As an indicator of the number of claims, we obtain *publn_claims_max_tls211*, which is the maximum number of claims registered on PATSTAT (tls211 table) in a citing INPADOC family. We do not simply rely on claims data from a single office such as from the EPO, since an application can be modified during its prosecution internationally. We also employ *invt_nr*, the maximum number of inventors in an application included in a citing INPADOC family, from PATSTAT (tls207). The size of international family, *family_size*, is a count variable of applications in different countries in a citing INPADOC family (tls211/219).
In addition to the variables above, which are used to test hypotheses directly, we define three variables to address self-selection of ISAs by applicants. The first two represent the potential of the applicant. The first of the two is *total_count*, which is the number of total applications that an applicant has made, taken from EEE-PPAT. The second one is *applicant_avg_cited*, which is the number of average forward citations that an applicant has received, calculated by PATSTAT (tls212) and EEE-PPAT. Both are supposed to represent the experience level of the applicant, and are used as instrument variables for instrumented PROBIT on the variable *ISA_CHANGED*. This binary variable *ISA_CHANGED* indicates that the U.S. and Japanese applicants choose the EPO as their ISA (the EPO can be chosen from the U.S. and Japanese applicants, but not vice versa). This information can be obtained for PCT applications on PATSTAT, since the citation table tls212 has a field on "citation origin" where "ISR" is shown for PCT applications. Since first application country (RO) in a family is available from tls201, switching from RO to a different ISA can be coded. The correlation coefficient between *ISA_CHANGED* and the dependent variable *found_in_ISR* is low at 0.0348.
Control variables for originating areas, which are *JP_app* and *US_app* (applications from Japan and the U.S., respectively), are used. Technology class is controlled by thirty-five WIPO technology classification dummies (results not shown for space reason).

Estimation results

The result shown in the Model 1 of Table 1 employs all samples from the triadic regions. As is evident from the negative sign for *JP_app* and *US_app*, the baseline ISA (EPO) is found to be advantaged in finding prior art at the time of ISR. The positive sign of *ISA_CHANGED* also indicates that prior art is easier to be identified at the time of ISR if applicants from the U.S. or Japan choose the EPO as their ISA (for which robustness is checked in Model 4 and 5). These are consistent with the EPO's good reputation from international applicants. H1 is supported from the positive sign of *euro_cited*. Likewise, H2, H3, H4 and H5 are all supported o this model, except that the number of inventors has an insignificant coefficient.
Model 2 uses applications from Japan only in order to examine the locality of knowledge in Japan. As is expected in H1, *jp_cited* has a positive and significant sign, whereas *us_cited* has negative and significant sign. Other variables show similar results with the Model 1 and are consistent with hypotheses, except *self* indicates the negative sign. Model 3 uses U.S.

applications only, and the results are just consistent with the hypotheses. Model 4 and 5 limit the citation data to non-self citations only for robustness checks, while employing two instrument variables on the variable *ISA_CHANGED*. For Japanese applications, the coefficient for *ISA_changed* lost the significance in the Model 4, suggesting that the advantage provided by the ISA change from JPO to EPO is due to the applicants' self-selection. However, this effect is not observed for the U.S. applications in the Model 5.

Table 1. PROBIT analyses on the probability of ISR coverage; dep. var.=*found_in_ISR.*

Model 4 and 5 use "*total_count*" and "*applicant_avg_cited*" as instruments for "ISA_CHANGED."

****<0.001 ***<0.01 **<0.05 Robust standard errors are in the parentheses (clustering on citing family).

Model & sample	Model 1 (all of triadic samples/ baseline=EP_app)	Model 2 (JP app only)	Model 3 (US app only)	Model 4 (JP app & non-self only)	Model 5(US app & non-self only)
method	Probit	Probit	Probit	IV Probit	IV Probit
euro_cited	0.1419984**** (0.0080393)	-0.031025 (0.0160179)	0.1776262**** (0.0120059)	0.0203394 (0.0174625)	0.148418**** (0.0253879)
us_cited	-0.0620007**** (0.0078305)	-0.3377195**** (0.0155267)	0.050351**** (0.0114757)	-0.2974986**** (0.0169034)	0.0777813**** (0.0159886)
jp_cited	0.0393056**** (0.0082601)	0.8054234**** (0.0151802)	-0.4295359**** (0.0121628)	0.8367819**** (0.0175193)	-0.3751166**** (0.0427623)
fam_cite_lag	0.0030127**** (0.000212)	0.0023379**** (0.0004175)	0.0046464**** (0.000329)	0.0005303 (0.0004425)	0.0026492**** (0.0005495)
self	0.2091817**** (0.0047187)	-0.1759722**** (0.0082345)	0.1123806**** (0.0076398)		
fwd_cite_of_the_cited	0.0000359**** (0.00000321)	-0.00000566 (0.00000781)	0.0000573**** (0.00000437)	-0.00000566 (0.00000799)	0.0000551**** (0.00000526)
IPC4_count	-0.0165033**** (0.0013614)	-0.0176023**** (0.002381)	-0.0215867**** (0.0022476)	-0.0170435**** (0.0026306)	0.0099131 (0.011092)
publn_claims_max_tls211	-0.0080901**** (0.0001942)	-0.0029271**** (0.0003468)	-0.0094453**** (0.0002733)	-0.0033284**** (0.0004149)	-0.0081833**** (0.0010323)
invt_nr	0.0000932 (0.0011831)	-0.0007108 (0.002112)	-0.0058672*** (0.0018111)	0.0008906 (0.0023144)	-0.0089979*** (0.0026535)
family_size	-0.006626**** (0.0007439)	-0.0142835**** (0.0021553)	-0.0053694**** (0.0011327)	-0.0091501*** (0.0032126)	-0.0138593**** (0.002496)
JP_app	-0.0667862**** (0.0069462)				
US_app	-0.2808785**** (0.0072769)				
ISA_CHANGED	0.3096426**** (0.0066579)	0.2758815**** (0.0169662)	0.380766**** (0.0074961)	0.0109491 (0.1314658)	1.35421**** (0.3121653)
Technology class dummies	included	included	included	included	included
n	1031127	325990	455830	264805	363328

Discussion and further development

Overall results are consistent with the hypotheses, suggesting that examiners (and searchers working for the PTOs) are bound by various kinds of "distances," including technological complexity of applications. These are intuitive, and are supported by the novel methodology for the first time. An interesting interpretation is that examiners (unlike inventors) are

required to find prior art by law, but that they are naturally bound by informational horizons they have. This has policy implications, since Patent Prosecution Highways (PPH) rely on outcomes from previous patent offices. Most prior studies using examiner citations do not incorporate these informational obstacles born by examiners, but they cannot be ignored. For example, prior studies on the difference of examination outcomes between patent offices (Jensen et al., 2005; Webster et al., 2007, 2014) do not explicitly consider them, but the cost of prior art search may affect the results. The results with instrument variables suggest the self-selection is working, but is evident for the Japanese samples only. Further scrutiny is needed.

Acknowledgments

This interim output is drawn from the collaborative project with Professor Setsuko Asami (Tokyo University of Science) and Professor Yoshimi Okada (Hitotsubashi University). The entire project is supported by RISTEX/JST. The comparison of search quality between PTOs at aggregated level was previously presented at the International Workshop on Patent System Design for Innovation at Hitotsubashi University (Wada and Asami, 2014) and at the 2014 Annual Conference of the Asia-Pacific Innovation Network. The idea of the probability of ISR coverage of this paper evolved out of the idea of aggregate ISR coverage ratio, which Professor Asami first thought of. The author also acknowledges the support from the MEXT/JSPS (Grant #22330122) for the analyses of citations and firm boundaries.

References

Alcacer, J., & Gittelman, M. (2006). Patent citations as a measure of knowledge flows: The influence of examiner citations. *The Review of Economics and Statistics*, *88*(4), 774-779.

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.

Jaffe, A., & Trajtenberg, M. (1999). International knowledge flows: evidence from patent citations. *Economics of Innovation and New Technology 8*, 105-136.

Jaffe, A., & Trajtenberg, M. & Henderson, R. (1993). Geographic localization of knowledge spillovers as evidenced by patent citations. *Quarterly Journal of Economics*, *108*, 577-598.

Jensen, P.H., Palangkaraya, A. & Webster, E. (2005). Disharmony in international patent office decisions. *Federal Circuit Bar Journal*, *15*, 679.

Magerman T, Grouwels J., Song X. & Van Looy B. (2009). Data Production Methods for Harmonized Patent Indicators: Patentee Name Harmonization." EUROSTAT Working Paper and Studies.

Peeters B., Song X., Callaert J., Grouwels J., & Van Looy B. (2009). Harmonizing harmonized patentee names: an exploratory assessment of top patentees. EUROSTAT working paper and Studies.

Wada, T., & Asami, S. (2014). Quality comparison of International Search Reports (ISRs) by selectable International Search Authorities (ISAs) under the Patent Cooperation Treaty (PCT) system, *a paper presentation at the 2014 International Workshop on Patent System Design for Innovation,* Hitotsubashi University, Tokyo, Japan.

Webster, E., Jensen, P. H. & Palangkaraya, A. (2014). Patent examination outcomes and the national treatment principle. *The RAND Journal of Economics*, *45*, 449–469.

Webster, E, Palangkaraya, A & Jensen, P.H. (2007). Characteristics of international patent application outcomes. *Economics Letters* 95, 362-368.

World Intellectual Property Organization. (2014). Patent Cooperation Treaty Yearly Review: The International Patent System. *WIPO Economics and Statistics Series*.

A Collective Reasoning on the Automotive Industry: A Patent Co-citation Analysis

Manuel Castriotta and Maria Chiara Di Guardo

{manuel.castriotta, diguardo}@unica.it,
University of Cagliari (Italy)

Abstract
While collective cognition has received increasing attention in the broader field of organization, academic research has largely overlooked its potential role on shaping innovation trajectories and technological change adaptation at a firm and industrial levels. Through a strategic lens and based on the patent bibliometrics and patent co-citation methods, we integrate and extend the cognition and technology strategy literatures by proposing an invention behavior map of leading companies and groups in the automotive industry. How collective cognition influence patent strategies? How economic trends impact on patent paths? Empirical evidence for these reasons is drawn from a longitudinal patent analysis quantitative approach of the period 1991-2013 considered overall and consequently subdivided into three sub periods of seven years each 1991-1997, 1998-2004, 2005-2013. About 443.000 patents, 1.108.356 citations and 1.234.623 co-citations of 49 automotive assignees were collected from Derwent Innovation Index (DII), the largest world patent and innovation database. Multi dimensional scaling and cluster analysis techniques are employed to detect embryonic cognition homogeneity measures and provide an overview of groups technology composition and companies innovation strategies trends. Finally, explorative findings are discussed below with suggestions about how they might be translated into managerial implications.

Conference Topic
Patent Analysis

Introduction
The empirical literature on technological regimes argues that firms within an industry behave in correlated ways because they share sources of information and technology (suppliers, universities, other industries), and perceive similar opportunities for innovation. The existence of a collective cognition shared by firms within a sector can also influence how inventions arise and how quickly and completely they diffuse, and can give us another key to better understand the collective failure of some industries as a result of surprisingly unexpected technological changes, or the innovation trajectories that have characterized some sectors. Yet, while collective cognition has received increasing attention in the broader field of organizational theory (Johnson & Hoopes, 2003; Nadkarni & Narayanan, 2007), research on innovation and patent strategies has been largely silent about the cognition's role (Kaplan, 2011, 2012; Kaplan & Tripsas, 2003, 2008) and empirical studies thus far have not questioned how industry boundaries truly define patent strategies and how economic trends impact on technological trajectories.

To take the first steps at going beyond these limitations and embryonically understand how industry structure and interaction among players can shape technological trajectories, we examine the case of the automotive sector from 1991 to 2013 and identify the dynamic evolution of patent paths among the principal actors in this sector. We chose the automotive sector for several reasons: first, the ability of firms to innovate is crucial to commanding a competitive advantage in this industry (Norhia & Garcia-Pont, 1991); second, all relevant players in this industry must routinely patent their innovations; and third, the automotive market is characterized by high entry barriers able to isolate new entrants and incumbents' dynamic noise.

In order to understand the phenomenon at stake, we analyze the evolution of the technological trajectory in the automotive sector by utilizing bibliometric information such as patent co-citations (Lai & Wu, 2005; Wang, Zhang & Xu, 2011). This approach displays a larger picture of the overall innovation structure and the patent linkages among players and groups' technology positioning, thereby shedding light on the patterns of patent strategies within an industry.

In total, a 21-year period, subdivided as three sets of years in seven-year time spans from 1991 to 1997, 1998 to 2004, and 2005 to 2013, are visualized. About 443.000 patents, 1.108.356 citations and 1.234.623 co-citations of 49 automotive assignees were collected from Derwent Innovation Index (DII), the largest world patent and innovation database. Multidimensional scaling and cluster analysis techniques are employed to detect the embryonic cognition homogeneity measures and to provide an overview of the groups' technology composition and companies' innovation strategy trends.

This study adds to the literature in multiple ways. First, it contributes to the patent literature showing the evolutionary patterns of patent strategies inside a specific industry using patent co-citation analysis. Second, it contributes to innovation literature by enhancing our understanding of how technological firms and group positioning evolve and are influenced by collective cognition. Third, it also contributes to the still-inadequate understanding of the drivers of patent strategies and innovation trajectories.

The paper is organized as follows. In section two, we describe the patent co-citation methodologies employed; in section 3, we present the bibliometric results and provide a graphical representation of firms' and groups' proximities performed by multidimensional scaling (MDS) and cluster analysis; in section 4, we discuss embryonic results and offer some conclusions;

Theoretical background

Bibliometrics and patent citation analysis

Patent citation analysis is an academic set of bibliometric methods directly derived from methodology that seeks to link patents in the same way that science references link papers. Papers and patents are both research instruments that adopt citation-count measurement systems (Narin, 1994). Moreover, in bibliometrics, the use of a citation approach for the assessment of similarity for the classification of documents is a mature methodology, and for this reason, it is feasible to apply the citation analysis of bibliometrics to patent analysis (Zhao & Guan, 2013).

Patent co-citation analysis

Co-citation analysis is a measure of the frequency of how many times A and B units are co-cited by third earlier units such as papers, authors, institutions, and in our study patents, inventors, or assignees (Lai & Wu, 2005; Wang et al., 2011). The assumption of co-citation analysis is that documents that are frequently cited together cover closely related subject matter (Small, 1973; Narin, 1994). In this vein, the co-cited frequency of patents can be used to assess the similarities or relatedness and to post evaluation and less-subjective unobtrusive patent maps and classification systems (Lai & Wu, 2005). In bibliometrics, it is used to assess document similarities in order to analyze the intellectual structure of science studies and identify cluster specialties and sub-fields (McCain, 1990; Di Guardo & Harrigan, 2012; Di Stefano, Gambardella & Verona, 2012).

Methodology

Sample and unit of analysis selection

Our analysis, following the bibliometric co-citation and patent co-citation methods prescriptions (McCain, 1990; Wang et al., 2011; Di Guardo & Harrigan, 2012) and in order to correctly select the unit of analysis started by tracing the history of most relevant M&As and alliances automotive industry milestones. This allow us to consequently identify in Derwent database the standard and non standard assignees codes for the overall and intermediate periods and correctly formulate compound Derwent Innovation Index and Derwent World Patent Index search queries (Wang et al., 2011). We retrieved assignees patent bibliometrics and assignees patent citation counts and finally co-citation frequencies. Operationally, the compilation of the raw co-citation matrix and its conversion to correlation matrix allow us to run multivariate analysis and consequently interpreting the findings. In the case of academic bibliometric studies, the unit of analysis may consist of scientific articles, authors and institutions (Small, 1973). Symmetrically, in the study of citation behavior in the patent analysis, the unit of analysis can be identified by single patents, inventors, institutions or assignees (Lai & Wu, 2005). Our research aims to show the strategic positioning and similarities between the leading automotive companies by displaying and then comparing the entire period of time with three different timespans. For these reasons we adopted assignees as unit of research.

Starting from the OICA 2013 report ranking, we selected the top 80 global companies in the automotive industry of manufacturers based on the number of commercial, passenger, and industrial vehicles produced. We examined the companies' websites and identified the number of brands for each company and its automotive groups. In the Derwent database, we checked individually for brands, single companies and groups, and the number of patents of the application date for the period 1991 to 2013. In this way, we divided the commercial brands by independent enterprises capable of producing technology. Then we looked back across the brands' histories, alliances, and M&As that occurred in the years between 1991 and 2013. In addition, in order to avoid the traditional limitations due to strategic and formal changes in companies and group structures, Derwent provides a comprehensive data set of joint ventures drawn up within industries in the period considered. From the operational point of view and following the correct search strategy proposed by Wang et al. (2011), we did a screening of all potential Derwent codes, including those with a different denomination than the main automotive group, related to joint ventures and M&As. In the research, we took into consideration 14 joint ventures formalized during the period among 18 companies.

Then, we launched an investigation of patent bibliometrics and identified the number of citations of the top 60 car manufacturers. Furthermore, in the hope of exploring the potential effects of the crisis in the strategic positioning of technology groups, we considered these in conjunction with the Asian crisis of 1997 - 98 and just before the start of the crisis of 2007–2008. Moreover, we took into account the M&A histories that showed that in these three periods, the most influential automotive group changes were concentrated. By analyzing the three periods, it was possible to visualize the structural change trends of automotive world industry. Finally, through the multidimensional scaling, a methodology that reduces the complexity and allows the matrices of proximity of certain objects to be studied (Mc Cain, 1990), we displayed the shape and measure the density of automotive sector conformation.

Discussion of results

Patent co-citation

The analysis of co-citations highlights the strategic positioning of the 49 major technological automotive companies in the global market in the period 1991 to 2013, 28 of the main groups in the periods 1991 to 1997 and 1998 to 2004, and finally the 34 major groups between 2005 and 2013. During the full period, the unit of analysis is the single automaker, while in the three time spans it is the automotive group through the extraction of aggregate data. The analysis of the complete map and the trends and changes in technology portfolios in the three time spans, considering the M&A histories and joint ventures, are discussed below through the results of multidimensional scaling and cluster analysis.

MDS and Cluster Analyses

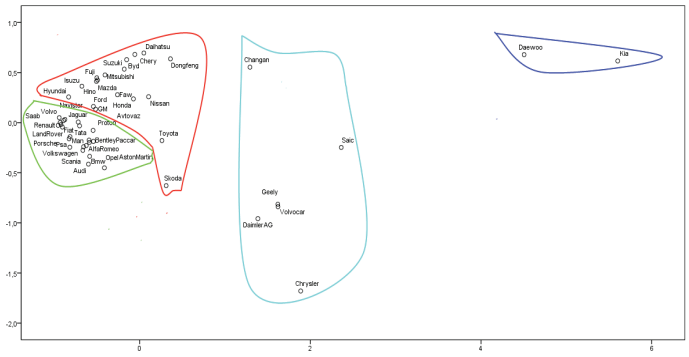


Figure 1. 1991-2013.

On the left of Figure 1 shows an area of high concentration and high technological similarities, while on the right, the distances among firms increase. In this scenario, cluster analysis clearly highlights four groups. The Japanese firms Toyota, Honda, and Nissan are the most central companies and belong to a larger international group comprised of Japanese, Chinese, Korean, and US companies. On the bottom left of the map, European manufacturers emerge, such as Volkswagen, Fiat, Porsche, Renault, BMW, PSA, and MAN, among which are India's Tata and the Soviet Avtovaz and the Malaysian Proton and its Lotus brand. Ford, GM, and Hyundai represent a technological bridge between the two areas. An important peculiarity of some company outliers such as Chrysler, Daimler AG, Geely, Volvo, and Chinese Saic and Dongfeng that belong to cluster 3 is seen, while peripheral positioning is occupied by Daewoo and Kia at the top right.

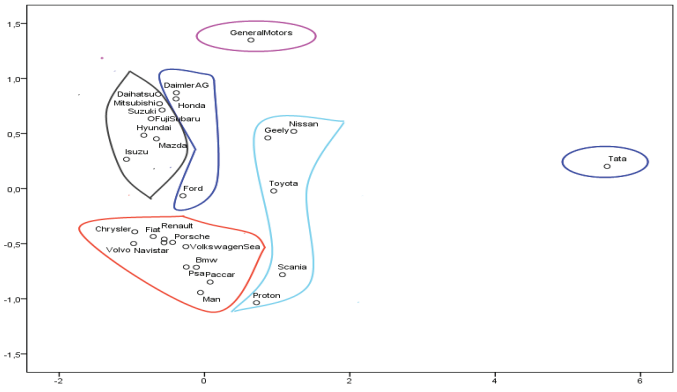


Figure 2. 1991-1997.

Figure 2 shows a major cognition concentration among firms, with the exception of the Indian company Tata on the right side. Ford, Toyota, and Renault are the major groups of centrality. Geely is the only Chinese enterprise present. Cluster analysis clearly shows six groups. General Motors is highly decentralized, a symptom of the uniqueness of its patent portfolio. Daimler and Hyundai are central, positioned in the two groups at the top along with the major Japanese companies, while at the bottom are MAN, Navistar, Volvo, and Paccar, which are all specialized in truck production, just below the European Union automakers. Interesting is the proximity of technology for Fiat and Chrysler, now belonging to the same group, and vice versa, the distance between Toyota and Daihatsu as separate companies at that time and since 1999 part of the same group. Of note is the proximity between Porsche and Volkswagen. Finally, the Volvo Group, at this stage not yet divided between truck and car production, is positioned at the left side near Navistar.

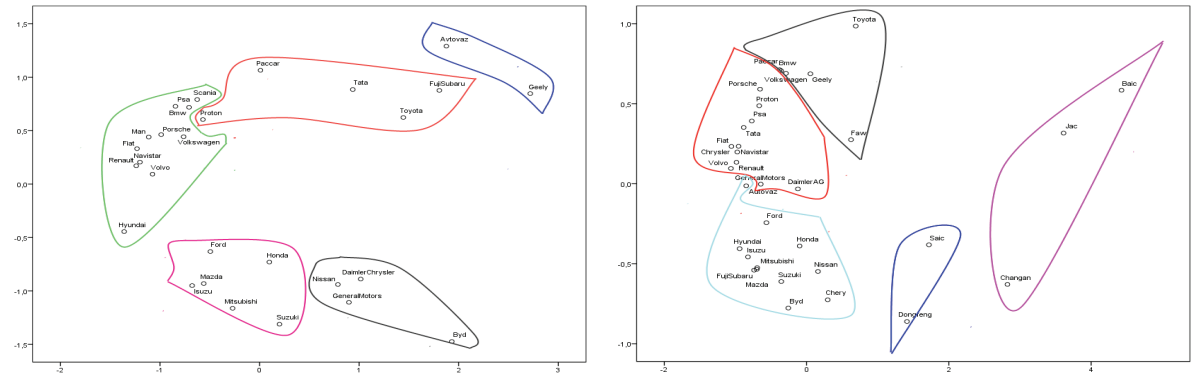


Figure 3. (a) 1998–2004. (b) 2005-2013.

Figure 3(a) transposes the effects of the Asian crisis of 1997-1998 and has a strong dispersion compared to the previous period's technology structures. The distances between companies are larger. To highlight the lack of a technological leader and a high level of technological heterogeneity, the central part of the map is empty.

Figure 3(b) includes the effects of the strong economic performance and global sales of the previous five years to have a stronger concentration symptomatic of technological proximity than in the previous period. During this period, Daimler AG, Ford, and GM occupy the most central locations on the map. General Motors, in particular, takes a decidedly opposite path in the three periods compared to Toyota. The American company tends to centralize its positioning technology, while Toyota tends to move within the confines of the map.

Conclusion and Limitations

This exploratory study increases the awareness of scholars by detecting and visualizing the cognitive structure, operationalized as companies' technological distances, of the automotive sector between 1991 and 2013. It reveals innovation similarities, technology positioning, and trends of assignees and groups, and makes it possible to hypothesize patent strategies and latent relationships among them. A contribution to the patent strategy and cognition literature has emerged on the basis of differences in positioning among companies and groups during the entire period and divided into time spans. In the overall map, this has emerged as some groups are composed of firms with heterogeneous positioning and consequently heterogeneous patent portfolios, while other groups have steadily increased over the years by acquiring high map closeness with companies with similar technological characteristics.

Second, the analysis of the three subdivided periods has highlighted how the level of similarity or distance among the groups, namely the collective cognition, changes continuously. The high concentration level that characterizes the first period is changed in the

second, which is more dispersed and where there are not central or technological leader groups. Yet the third one returns to a concentration level similar to the first period. Such behavior of the map, if considered in relation to the economic performance of the production and sales of the industry, reveals how, in times of crisis, companies tend to look for a heterogeneous technology portfolio to obtain competitive advantages, while in positive economic periods, conformity tends to prevail. It is as if the collective cognition profoundly affects the technology positioning and behavior of firms at the expense of objective assessments of patent strategy decisions. Third, research has highlighted significant strategic differences in positioning in the various periods in which such central enterprises move to the suburbs and vice versa, and some change their technology cluster membership by moving into another and finally emerge or disappear because of a failure or because of an M&A.

Fourth, an explorative contribution originates from the evaluative study of the groups' conformation in terms of brands and partnership formal contracts. In fact, it opens new horizons to researchers who want to analyze the impact of M&As or JVs on technological map positioning and, for example, in Foreign Direct Investments (FDI) and technology strategy literature. Finally, explorative findings of this study might be translated into managerial implications from the point of view of the companies strategic positioning planning. In fact, by detecting the heterogeneous technologies adoption (displayed by the more distant nodes in MDS), manager can potentially create innovative patent recombination strategies and consciously determine innovative future technological positioning scenarios.

References

Di Guardo, M. C., & Harrigan, K. R. (2012). Mapping research on strategic alliances and innovation: a co-citation analysis. *The Journal of Technology Transfer*, *37*(6), 789-811.

Di Stefano, G., Gambardella, A., & Verona, G. (2012). Technology push and demand pull perspectives in innovation studies: Current findings and future research directions. *Research Policy*, 41(8), 1283-1295.

Johnson, D. R., & Hoopes, D. G. (2003). Managerial cognition, sunk costs, and the evolution of industry structure. *Strategic Management Journal*, *24*(10), 1057-1068.

Kaplan, D. M. (2012). How to demarcate the boundaries of cognition. *Biology & Philosophy*, *27*(4), 545-570.

Kaplan, S. (2011). Research in cognition and strategy: reflections on two decades of progress and a look to the future. *Journal of Management Studies*, *48*(3), 665-695.

Kaplan, S., & Tripsas, M. (2003). Thinking about technology: understanding the role of cognition and technical change. Division of Research, Harvard Business School.

Kaplan, S., & Tripsas, M. (2008). Thinking about technology: Applying a cognitive lens to technical change. *Research Policy*, *37*(5), 790-805.

Lai, K. K., & Wu, S. J. (2005). Using the patent co-citation approach to establish a new patent classification system. *Information Processing & Management, 2*, 313-330.

McCain, K. W. (1990). Mapping authors in intellectual space: A technical overview. *Journal of the American society for information science*, *41*(6), 433-443.

Nadkarni, S., & Narayanan, V. K. (2007). Strategic schemas, strategic flexibility, and firm performance: the moderating role of industry clockspeed. *Strategic management journal*, *28*(3), 243-270.

Narin, F. (1994). Patent bibliometrics. *Scientometrics*, 30(1), 147-155.

Nohria, N., & Garcia-Pont, C. (1991). Global strategic linkages and industry structure. *Strategic management journal*, *12*(S1), 105-124.

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*, *24*(4), 265-269.

Wang, X., Zhang, X., & Xu, S. (2011). Patent co-citation networks of Fortune 500 companies. *Scientometrics*, *88*(3), 761-770.

Zhao, Q., & Guan, J. (2013). Love dynamics between science and technology: some evidences in nanoscience and nanotechnology. *Scientometrics*, *94*(1), 113-132.

Statistical Study of Patents Filed in Global Nano Photonic Technology

Zhang Huijing[1], Zhong Yongheng and Jiang Hong

[1] *zhanghj@mail.whlib.ac.cn*

Wuhan Documentation and Information Centre, Chinese Academy of Sciences, West 25, Xiaohongshan, Wuhan, Hubei, PR, 430071 (China)

Key words

nano photovoltaic technology (NPT); patent analysis; review; photoelectron device; semiconductor material; industry layout

Introduction

As one of leading core technology in the 21th century, nano photonic technology (NPT) is highly interdisciplinary, involving physics, chemistry, biology, materials science, and the full range of the engineering disciplines (Picraux, 2014). NPT is a study of the interaction of electrons and photons and its components in nano structure based on the great development and popularization of nanometre semiconductor materials (Liu, 2005). In 2011, NPT was identified as one of Key Enabling Technologies (KETs) for its vital role in strengthening Europe's industrial and innovation capacity (European Commission, 2011). It is widely used in telecommunications, optical interconnects, display, lighting, photovoltaic, sensors, data storage, imaging, and testing, etc (AIRI/Nanotec IT, 2008).

Patent analysis, which involves statistical, analytical, and comparative methods for examining information in patent documents, has been widely applied in studies examining R&D capacity, technological fields, industrial departments, and company levels (Pavit, 1988). Careful analysis of NPT-related patents can assist in elucidating technological details and relationships, identifying business trends, inspiring novel industrial solutions, and developing investment policies. Therefore, this study performed a statistical analysis of patent data to explore the technological developments of NPT. The technology life cycle and regional distribution of the patents were studied, and the top ten patent assignees were also explored.

Methodology

The searching for NPT patents from the Derwent World Patent Index (DII) database, keywords search were performed for the term appearing in titles, abstracts, or claims. The search strategy of DII database based on NPT was as follows: TS=(((solar or photovoltaic or "optoelectronic integrated device" or OEIC or "optic switch" or "holographic memory" or "light amplifier" or "optical amplifier" or ROADM or "optical add-drop multiplexer" or "optoelectronic display") and nano) or (optoelect* and (semiconductor or GaAs or "gallium arsenide") and nano) or (("quantum well" or "quantum wire" or "quantum dot") and (laser or "photoelectric effect")) or "micronano laser" or "nano laser" or Nanophot* or "Nanowire laser" or "Uv nm laser" or "microcavity laser" or (nano same LED) or (nano same "light emitting diode")). After querying, filtering, and organizing the search results, 8168 NTP-related patents were obtained on December 12, 2014, and the data were analyzed using Thomson data analyzer (TDA).

Results and discussion

Figure 1 showed the evolution of the number of patents relative to the assignees, which is a typical value for exploring the technology life cycle base on patent data. It was showed that the number of patents and assignees increased gradually before 2000, indicating that the technology life cycle was in the introductory stage. This trend implied that few manufactures and institutions were investing in the R&D of NPT before 2000. By contrast, the number of patents and assignees increased rapidly after 2000, particularly during the 2007-2013 periods, indicating that the technology had entered the growth stage. Specifically, the number of patents (assignees) increased from 378 (558) in 2007 to 1006 (843) in 2013.

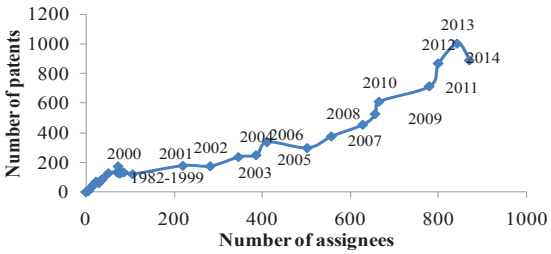


Figure 1. Technology life cycle

Figure 2 showed the number of patents filed in various countries/offices, as well as the trend of the number of patent applications. China (CN), Japan (JP), United State (US), WIPO (WO), and Korea (KR) were the top five countries/offices, with the number of patent applications of 2133, 1964, 1946, 970 and 656. The number of patent applications filed in CN was the highest, indicating that the NPT market in CN might offer the most potential for future development. Compared with other countries, the filing of NPT-related patents commenced only recently in CN, although the number of patent applications increased markedly in 2004-2014.

Moreover, the NPT-related patents were filed earliest in US and WO, and the number of patent applications of these two countries grew rapidly since the beginning of 2004.

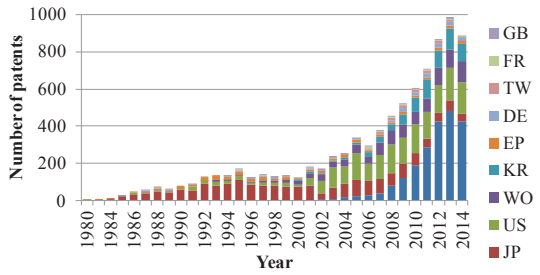


Figure 2. Number of patents and its evolution by country/office. The initialisms "WO" and "EP" indicate that the patent was filed in the WIPO and EPO, respectively.

Table 1 showed a summary of the top ten patent assignees. It was found that all of top ten patent assignees were from JP except *Semiconductors Institute of Chinese Academy of Sciences* and *Samsung Electronics Company, Limited*. In addition, the JP assignees were all companies, and these JP companies had already manufactured commercial NPT products. Furthermore, the JP and KR assignees were filed their patents in many countries/offices for the global layout of NPT. By contrast, *Semiconductors Institute of Chinese Academy of Sciences* filed patents only in CN.

Table 1. Top 10 patent assignees.

Assignee (nationality)	No. of patents	No. of application countries	Times cited (average)
NEC Corporation (JP)	280	4	425 (1.5)
Mitsubishi Denki K.K. (JP)	188	7	402 (2.1)
Fujitsu Limited (JP)	179	5	210 (1.2)
Sharp KK (JP)	170	6	430 (2.5)
Hitachi Limited (JP)	156	4	187 (1.2)
Samsung Electronics Company, Limited (KR)	153	6	137 (0.9)
Semiconductors Institute of Chinese Academy of Sciences (CN)	143	1	71 (0.5)
Furukawa Electric Company, Limited (JP)	138	6	423 (3.1)
Nippon Telegraph & Telephone Corporation (JP)	132	3	30 (0.2)
Matsushita Denki Sangyo KK (JP)	115	5	286 (2.5)

Conclusion

This study analyzed patent data to explore the technological developments of NTP. After querying, filtering, and organizing the search results, this study analyzed 8168 NTP-related patents. The primary findings of this study were detailed as follows.

(1) Based on the analysis results, the technology life-cycle status of the NPT is currently in the growth stage, indicating that many products were sufficiently developed for commercialization.

(2) US assignees were the most prominent assignees, although the most patent applications were filed in CN, indicating that the market for NPT in CN might offer the most potential for future development.

(3) All of the top ten assignees were from JP, KR, or CN. The JP and KR assignees were all companies, and the assignees were filed their patents in many countries/offices for the global layout of NPT and products. By contrast, *Semiconductors Institute of Chinese Academy of Sciences* is academic institution and filed patents only in CN.

Future studies should consider evaluating the current state of NPT developments in a specific field to identify application areas for new patents.

Acknowledgments

We deeply appreciate the financial supports to this research from Youth Innovation Promotion Association of Chinese academy of sciences.

References

AIRI/Nanotec IT (2008). Roadmmaps at 2015 on nanotechnology application in the sectors of: materials, health & medical systems, energy. Retrieved December 12, 2014 from: http://www.iva.se/upload/Verksamhet/Projekt/Nano/internetionellt/EU%20Nano%20Roadmaps%20SYNTHESIS.pdf.

European Commission (2011). High-Level Expert Group on Key Enabling Technologies. Retrieved December 12, 2014 from: http://ec.europa.eu/enterprise/sectors/ict/key_technologies/kets_high_level_group_en.htm.

Liu W. L. (2005). New advancement and developmental trend of nanometer optoelectronic devices. Sensor World, 11, 6-9.

Pavitt K. (1988). Uses and abuses of patent statistics. In Van Raan AFJ (Ed.), Handbook of quantitative studies of science and technology (pp: 509-536). Amsterdam: Elsevier Press.

Picraux S.T. (2014). Nanotechnology. Retrieved December 12, 2014 from: http://global.britannica.com/EBchecked/topic/962484/nanotechnology.

A SAO-based Approach for Technology Evolution Analysis Using Patent Information: A Case Study on Graphene Sensor

Zhengyin Hu[1,2] and Shu Fang[1]

huzy@clas.ac.cn
[1]Chengdu Documentation and Information Center, Chinese Academy of Sciences,
No.16, Nan'erduan, Yihuan Road, Chengdu (China)
[2]University of Chinese Academy of Sciences,
No.19A Yuquan Road, Beijing (China)

[1] *fangsh@clas.ac.cn*
[1]Chengdu Documentation and Information Center, Chinese Academy of Sciences,
No.16, Nan'erduan, Yihuan Road, Chengdu (China)

Introduction

The Subject-Action-Object (SAO) structures are composed of Subject (noun phrase), Action (verb phrase) and Object (noun phrase), which can represent technology information with more details in a simple manner and have been widely applied in patent text mining (Cascini, Lueehesi, & Rissone, 2001; Sungchul et al., 2012; Zhang et al., 2014a). This paper presents an approach for technology evolution analysis based on SAO. SAO structures are extracted and cleaned from patent text. The technology information of patents such as problems, solutions, functions and effects are stated by SAO. By calculating the distributions of problems over solution groups, a technology evolution map of problems can be drawn. Graphene sensor patents are selected as a case study.

Methodology

Extracting SAO Structures

After collecting patents, some national language processing (NLP) tools are used to extract raw SAO structures from patent text fields. Normally, the fields such as "Title" and "Abstract" are precise and meaningful for NLP (Sungchul et al., 2012).

Cleaning SAO Structures

The number of raw SAO structures is huge and they need to be cleaned. Text mining tools and domain thesauri are used to carry out Subject and Object cleaning by following a term clumping framework (Zhang, et al., 2014b). The verb phrases of Action are normalized and categorized by experts.

Tagging SAO Structures

According to a classification model learned from a training data, the cleaned SAO structures are tagged with 4 kinds of labels of *problem*, *solution*, *function* and *effect*.

Clustering SAO of Solution

After tagging the semantic type of each SAO, those with *solution* label are clustered into different *solution* groups. Each solution group with similar SAO can be considered as a *solution* topic.

Drawing technology evolution map of problems

Kim, Suh and Park (2008) approached a method that can be used to draw technology evolution map of keywords by calculating the distributions of keywords over the keyword cluster groups. We draw technology evolution map of problems based on Kim, Suh and Park's (2008) research. Firstly, we calculate the distributions of problems over the *solution* groups. If the co-occurrence frequency of two problems is above a threshold, we draw a directed line segment between them to show their relevance. Then the occurrence frequency of each problem in *solution* groups is counted. Finally, by adding the earliest filling date of each problem, a technology evolution map of problems with horizontal axis of timeline and vertical axis of frequency can be drawn.

Case Study

Extracting SAO Structures

We selected Derwent Innovations Index (DII) as data source and invited experts to determine the patent retrieval strategy for graphene sensor patents. After eliminating irrelevant patents, we got 196 patents. We extracted raw SAO from the "Title" and "Abstract" fields and got 4,823 raw SAO structures using an NLP tool named ReVerb (Anthony, Stephen & Oren, 2011).

Cleaning SAO Structures

We cleaned Subject and Object by using a commercial text mining tool, VantagePoint (Nils, 2011) and domain thesauri. We followed the term clumping framework to clean them, which includes general cleaning, terms pruning and terms

consolidating processes. After term clumping, we got 628 terms of Subject and Object. We normalized and categorized the verb phrases of Action based on a rule table made by experts. After the cleaning steps, we got 2250 SAO structures.

Tagging SAO Structures

We chose 167 SAO structures from 20 patents as a training set. We picked up Subject, Action as the classification features and C4.5 decision tree as the classifying algorithm to build a classification model which helps to categorize SAO to 4 classes of *problem*, *solution*, *function* and *effect*. Among the classified SAO structures, there are 208 tagged with *problem* label, 746 with *solution* label, 824 with *function* label and 472 with *effect* label. A sample of SAO is shown in table 1.

Clustering SAO of Solution

We clustered the SAO structures with *solution* label into *solution* groups using *k*-means algorithm. By comparing the cluster results, we set the *k*-value 20 and got 20 *solution* groups.

Drawing technology evolution map of problems

By calculating the distributions of problems over each *solution* group, a technology evolution map of problems in graphene sensor patents was drawn. A part of the map is shown in Figure 1.

Table 1. A sample of SAO after tagging.

Type	Subject	Action	Object
Problem	method	synthetize	graphene oxide
Solution	method	use	ultrasonic oscillation process
Solution	graphite powder	mixed with	sodium nitrate
Function	graphene oxide	used for	thin film transistor

Conclusions

The technologies in the upper left corner of Figure 1 appeared in many different *solution* groups and were applied for patents in earlier time, which can be considered as the basic problems in graphene sensor, such as ***producing carbon nanotube***, ***synthetizing graphene oxide***, etc. The technologies in the lower right corner of Figure 1 appeared in fewer *solution* groups and were applied for patents lately, which can be considered as the latest technologies or emerging technologies, such as ***manufacturing sensor array***, ***detecting nucleic acid***, etc**.**

We can draw a technology evolution map of solution, function or effect by following a similar process. The separate technology evolution maps of problem, solution, function and effect can be combined to a more comprehensive technology evolution map of graphene sensor. This study is ongoing.

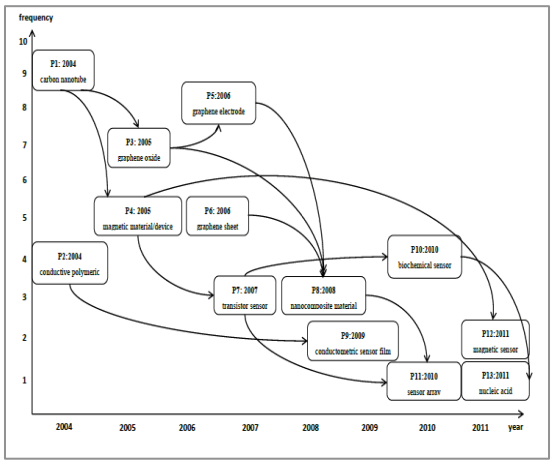


Figure 1. A part of technologies evolution map of problems in graphene sensor patents.

References

Anthony, F., Stephen, S., & Oren E. (2011). *Identifying Relations for Open Information Extraction*. Retrieved March 2, 2014 from: http://ai.cs.washington.edu/www/media/papers/reverb.pdf.

Cascini, G., Lucchhesi, D. & Rissone, P. (2001). Automatic patents functional analysis through semantic processing. *The 12th ADM International Conference*. Rimini, Italy.

Nils N. (2011). *VantagePoint*. Retrieved April 24, 2015 from: https://www.thevantagepoint.com/data/documents/VP%20INTRO%202011.pdf.

Sungchul, C., Hyunseok, P., Dongwoo, K., Lee, J.Y., & Kim, K. (2012). An SAO-based text mining approach to building a technology tree for technology planning. *Expert Systems with Application*, *39*, 11443-11455.

Kim, Y.G., Suh, J.H. & Park, S.C. (2008). Visualization of patent analysis for emerging technology. *Expert Systems with Applications*, *34*, 1804–1812.

Zhang, Y., Porter, A. L., Hu, Z., Guo, N., & Newman, N.C. (2014b). "Term clumping" for technical intelligence: A case study on dye-sensitized solar cells. *Technological Forecasting and Social Change*, *85*, 26-39.

Zhang, Y., Zhou, X., Porter, A. L., & Gomila, J. (2014a). How to combine term clumping and technology roadmapping for newly emerging science & technology competitive intelligence: The semantic TRIZ tool and case study. *Scientometrics, 101*(2), 1375-1389.

Prediction of Potential Market Value Using Patent Citation Index

HeeChel Kim[1,2], Hong-Woo Chun[2], Byoung-Youl Coh[2]

{kim, hw.chun, cohby}@kisti.re.kr

[1]University of Science and Technology, 305-350, 217 Gajeong-ro, Yuseong-gu, Deajeon(South Korea)

[2]Korea Institute of Science and Technology Information, Dept. Of Technology Intelligence Research, 130-741, 66 Hoegiro, Dongdaemun-gu, Seoul (South Korea)

Introduction

Patent statistics have frequently been used as both technological and economic indicators, however, in order to fully utilize patent data in economic analyses, we must link patents to economic activity at a level of industry or product.

Many previous pieces of research showed the effectiveness of patents citation index (PCI), containing annual citation information, on economic indicators of respective firms. Hall et al. (2005) have studied the relation between a market value and PCI using the Tobin's q approach, and Patel and Ward (2011) have compared the stock market value of firms with the patent citation using the event study methodologies. Both studies showed that Patent statistics can be effectively used to micro-level economic analyses and the increase of PCI has the positive effect on the corresponding market value.

Meanwhile, our study aims to prove the effectiveness of PCI on the economic value of industry, so-called Meso-level study and, in this case, it is essential to develop technology-industry concordance method.

Method

The correlation analysis between Potential Market value (PMV) and PCI for the respective industry is carried out in three stages.

(1) Data concordance process. The market data was collected from Annual Survey of Manufactures (ASM) [1] in the US Census Bureau (http://www.census.gov) and PCI [2] data was collected from the patent set registered USPTO.

Next, we created an annual concordance matrix of IPC (international patent classification) 4-digit to NAICS (North American industry classification system) 6-digit (rev.2002, 2007, and 2012) by Algorithmic Links with Probabilities (ALP), ALP (Lybbert & Zolas, 2013), concordance method of the WIPO (http://www.wipo.int/). ALP is the most up-to-date method compared with those of YTC (Kortum & Putnam, 1997), OECD (Johnson, 2002) and DG (Schmoch et al., 2003).

Each IPC 4-digit is connected to multiple NAICS 6-digit probabilistically via a text mining-based matching rule.

PMV was calculated by model 1 as follows, and consequently, 593 annual pairs of PMV-PCI for each IPC were generated.

$$PMV_{ij} = \frac{\sum_{k=1}^{476} a_{ijk} \times b_{ik}}{\sum_{i=1}^{593} \sum_{k=1}^{476} a_{ijk} \times b_{ik}} \times \sum_{k=1}^{476} b_{ik} \quad \text{..... Model 1.}$$

a = Probability of IPC 4-digit to NAICS 6-digit
b = Value of shipment by NAICS in ASM
i = Year (2002 to 2013)
j = IPC 4-digit code (A01G, A01H, …, H05K)
k = NAICS 6-digit code (311111, 311119, …, 339999)

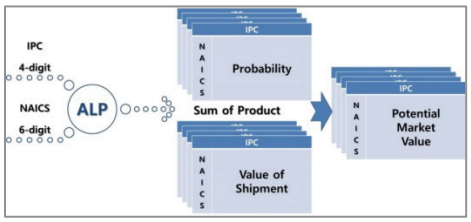


Figure 1. Process of IPC-NAICS Concordance and PMV Calculation.

(2) Statistical correlation analyses for all industry fields. We performed a statistical correlation analysis between the annual incremental of PMV and PCI. We used the Spearman's rho correlation analysis, a nonparametric correlation analysis algorithm, useful to calculate the correlation between the ranked variables (IBM, http://k:5172/help/index.jsp?topic=/com.ibm.spss.statistics.tut/introtut2.htm).

(3) Statistical correlation analyses for 4 major industry fields. The correlation analyses between the annual incremental of PMV and PCI for 4 major industry fields - electrical engineering, instruments, chemistry, and mechanical engineering – were also performed.

Result

Figure 2 shows annual trends of PMV, PCI, and Patent registered. All kinds of variables are trending upward in an accelerating degree.

[1]ASM is estimated sample statistics issued annually for more than one people employees firms in the manufacturing sector. ASM is classified industries by NAICS. In this study, using field of the value of shipment at the 2004 and 2006 edition of ASM that follow the revised NAICS 04 and 2008 to 2011 edition of ASM that follow the revised NAICS 07.

[2]PCI data was used granted patent of USPTO. During the year of from 2002 to 2013.

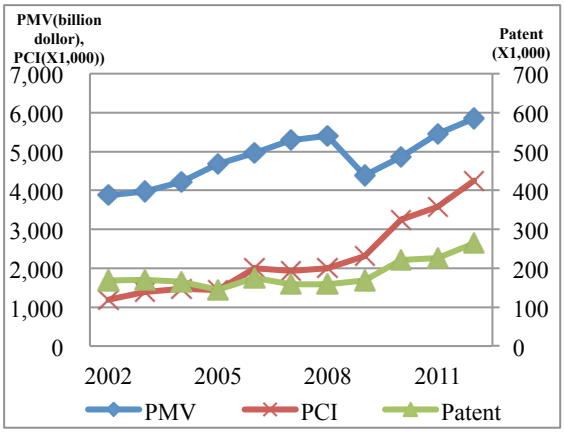


Figure 2. Structure of PMV, PCI and Patent.

PMV of each IPC

Table 1 shows the result of the PMV of each IPC calculated from model 1. It has a significant meaning that a set of patents can be expressed to market value.

Table 1. PMV (unit: million US$).

No.	IPC	2002	2003	…	2013
1	A01G	282	301	…	229
2	A01H	3,057	3,831	…	15,227
...	…	…	…	…	…
593	H05K	6,556	6,166	…	5,055

Correlation Analyses

In the analysis results over the entire industry fields (Table 2), we could find out that significance of correlation and direction varies depending on the Lagging time (differences in data collection year between PMV and PCI). It has a relatively weak positive correlation when the lagging time is 0, meanwhile, it showed relatively strong negative correlation when the lagging time is "PCI+1" – the data collection year for PCI is one year after to that of PMV - . And in case of the lagging time of "PCI-1", it has relatively strong positive correlation, which reveals patent citation activity's positive relation to the corresponding market value "one year later".

Table 2. Results of PMV-PCI rate's correlation analyses (all fields, **significance level 0.01).

Lagging time(year)	Correlation coefficient	p-value (two-tailed)	N
PCI-1	0.136**	0.000	5337
0	0.093**	0.000	5930
PCI+1	-0.323**	0.000	5337

The analyses results of 4 major industry fields showed similar tendencies to all-field-analysis except electrical engineering field.

Table 3. Results of PMV-PCI rate's correlation analyses (4 major fields, **significance level 0.01).

Field	Lagging time(year)	Correlation coefficient	p-value (two-tailed)
Electronic	PCI-1	-0.013	0.747
	0	0.143**	0.000
	PCI+1	-0.513**	0.000
Instrument	PCI-1	0.209**	0.000
	0	0.011	0.795
	PCI+1	-0.360**	0.000
Chemistry	PCI-1	0.180**	0.000
	0	0.022	0.434
	PCI+1	-0.265**	0.000
Mechanic	PCI-1	0.167**	0.000
	0	0.123**	0.000
	PCI+1	-0.266**	0.000

Conclusion

In this research, we made a systematic way for describing the technological impact on industry sector by using some indices, which has a significant meaning that a set of patents can be expressed to market value. We also had confirmed the potential of PCI to predict PMV of the industry. Experimental results showed that PMV in all industry fields was related by the corresponding field's patent-citation activity in one year before or after. After this work, we will deal with enhanced concordance approach to find out relationships between IPC 7-digit and NAICS 7-digit. Also, the self-citation ratio of patent-citation activity may affect economic activity at a level of industry or product, which is now on a study.

References

Hall, B. H., et al. (2005). Market value and patent citations. *RAND Journal of Economics*, 16-38.

Johnson, D., March (2002). The OECD Technology Concordance (OTC): Patents by Industry of Manufacture and Sector of Use, OECD Science, Technology and Industry Working Papers.

Kortum, S. & Putnam, J. (1997). Assigning patents to industries: tests of the Yale technology concordance. *Economic Systems Research*, *9*(2), 161-176.

Lybbert, T.J. & Zolas, N.J. (2014). Getting patents and economic data to speak to each other: An 'algorithmic links with probabilities' approach for joint analyses of patenting and economic activity. *Research Policy*, *43*(3), 530-542.

Patel, D. & Ward, M.R. (2011). Using patent citation patterns to infer innovation market competition. *Research Policy*, *40*(6), 886-894.

Schmoch, U., Laville, F., Patel, P., & Frietsch, R. (2003). Linking technology areas to industrial sectors. *Final Report to the European Commission, DG Research*, *1*.

Knowledge Flows and Delays in the Pharmaceutical Innovation System

Mari Jibu[1], Yoshiyuki Osabe[2], and Katy Börner[3]
[1] OECD, Paris (France) and Japan Science and Technology Agency, Tokyo (Japan)
[2]Japan Patent Office, Ministry of Economy, Trade and Industry, Tokyo (Japan)

[3] *katy@indiana.edu,*
CNS, ILS, SOIC & IUNI, Indiana University, Bloomington, IN (USA)

Introduction

This paper presents an analysis of knowledge flows in the pharmaceutical innovation process. Backward citations, citations to non-patent literature (NPL), and forward citations that link patents, scientific publications, and pharmaceutical pipelines data on drug developments are analyzed and visualized to provide a more holistic understanding. Results show that patents linked to drugs tend to be technically specialized when compared to patents without linkages to drugs. Moreover, patents linked to drugs tend to cite older patents and scientific publications and impact wider technological and scientific fields than pharmaceutical patents not linked to drugs.

Diverse studies have been conducted to study the origin, trajectory, and destination of knowledge flows and the delays in the science and technology system. Patents and citations between patents and to non-patent literature (NPL) are analyzed to understand knowledge spillovers (Lukach & Plasmans, 2002) or to measure patent quality (Squicciarni et al., 2013). The OECD Science, Technology and Industry Scoreboard 2013 (OECD, 2013) uses comprehensive and up-to-date data to report on knowledge flows via collaboration networks (e.g., derived from co-authored publications and co-inventors on patents), international migration of researchers (e.g., estimated from changes in author's addresses on publications), but also flows of royalty and license fees for technologies. Recently, the OECD introduced a new indicator, called "Patent-Science Link," that aims to measure knowledge flows between the science base and the innovation system (OECD, 2013). According to this new indicator, patented pharmaceutical inventions account for the majority of citations made from patents to scientific publications. That is, the distance between the science base and the innovation system is much closer in pharmaceutical fields than it is in other technological fields. Pharmaceutical innovation is particularly important for drug discovery, as research and development (R&D) costs are huge and major challenges exist for arriving at cost-effective new drugs. In fact, there is a steady decrease in R&D productivity over the last number of years (Booth & Zemmel, 2004).

The structure of the paper is as follows: The next Section details data acquisition and preparation. This is followed by a description of the methodology and results. The paper concludes with a discussion of key insights and their comparison to prior work.

Data Acquisition and Preparation

Five datasets by Thomson Reuters covering 1981 to 2011 are used in this analysis. (1) Publication data from the *Web of Science (WoS)* database. (2) Patent data from the *Derwent World Patents Index (DWPI)* and associated citations from the (3) *Derwent Patents Citation Index (DPCI)*. (4) Linkages between publications and patents come from the *WoS-DPCI Linktable* computed by Thomson Reuters and JST that provides information on backward citations from patents and to the non-patent literature (NPL), i.e., scholarly publications, derived from the *DPCI*. (5) Drug pipeline data was retrieved from the *Cortellis for Competitive Intelligence* database including detailed information of exactly drugs a patent is associated with. Data was compiled on December 11, 2013.

Interested to identify patents and their linkages to the NPL in pharmaceutical fields, we extracted all 833,376 patents with the International Patent Classification (IPC) code "A61P: Specific therapeutic activity of chemical compounds or medicinal preparations" from the *DWPI* with their citations from *DPCI*, called "Pharma_Patents." Then, we extracted 57,800 patents linked to pipeline data from the *Cortellis for Competitive Intelligence* database, called "Drug_Patents." Next, the *Drug-Patents* were subtracted from the *A61P-Patents* resulting in a dataset of 325,576 "Non-Drug Pharma Patents" that have the A61P code but are not linked to drugs.

Finally, all 115,252 NPL for *Drug_Patents (DP)* and 718,269 *Non-Drug_Pharma_Patents (NDPP)* were retrieved using the *WoS-DPCI Linktable*.

Methodology

Four metrics were computed: (1) *citation lag*; (2) *generality index* computing the diversity of patents that are cited by a given focal patent as well as the diversity of patents that are citing the focal patent;

(3) *subject index*, a new indicator based on the generality index but computed for NPL; (4) *patent scope*, often associated with the technological and economic value of patents with broad scope patents having a higher value (Lerner, 1994).

Results

Using the four metrics, a number of novel results can be computed.

Technology Delays: Citation Lag

Comparing citation lag data for *DP* and *NDPP* reveals the temporal dynamics of knowledge flows. Table 1 shows that forward citations from *NDPP* come from patents that were published on average 2.17 years later while *DP* are cited faster—after 1.89 years on average. Backward citations from *NDPP* go to patents that were published on average 3.4 years earlier and they go to much more recent NPL—published only 1.69 years earlier on average. Interestingly, *DP* cite older works than *NDPP*: Cited patents are 5.64 years old and cited NPL are 2.5 years old on average. All values are statistically significant at the 1% level. In sum, they show that *DP* cover larger temporal ranges and are cited more quickly than *NDPP*.

Table 1. Forward and Backward Citation Lags.

	NDPP	DP
Forward Cites by Patents	2.17	1.89
Backward Cites to Patents	3.40	5.64
Backward Cites to NPL	1.69	2.50

Technology Diversity: Generality & Subject Index

The generality index was calculated for 4- and 6-digit IPCs for forward and backward citations for *NDPP* and *DP*, see Table 2. *DPs* have higher generality index and subject index than *NDPP*. That is, on average, *DP* draw on more diverse technology "base knowledge" and are cited by a more diverse set of patents that have more varied IPCs. All values are statistically significant at the 1% level.

Table 2. Generality Index for Forward Citations (FC) and Backward Citations (BC).

		NDPP	DP
Generality Index (4-Digits)	FC	0.36	0.37
	BC	0.40	0.54
Generality Index (6-Digits)	FC	0.46	0.50
	BC	0.52	0.73
Subject Index	BC to NPL	0.22	0.28

Technology Value: Scope

The patent scope was computed for *NDPP* and *DP*, see Table 3. The scope of *DP* is lower than that of *NDPP*. This is unexpected as patents linked to drugs are presumably more valuable than those not linked to drugs.

Table 3. Scope.

	NDPP	DP
Scope (4-Digits)	0.13	0.11
Scope (6 Digits)	0.16	0.15

Conclusions

This paper compared and contrasted patents that are linked or not linked to drugs to understand knowledge flows and delays in pharmaceutical innovation. The results indicate that *Drug_Patents* draw from a more diverse set of technologies and are cited more widely across the technology landscape. However, they tend to be more technically specialized (lower scope) than *Non-Drug_Pharma_Patents*. Concerning citation lag, *Drug_Patents* tend to refer to older patents and scientific publications and are cited faster than *Non-Drug_Pharma_Patents*.

In our prior work, we introduced new drug-patent indicators for identifying patents related with pharmaceutical entities' R&D progress (Jibu & Osabe, 2014) and that IPC count, forward citations, and citations to NPL are efficient drug-patent-indicators. The work presented here is novel is that it shows that citation lags and the generality of backward citations are statically significantly different for *Non-Drug_Pharma_Patents* and *Drug_Patents*.

Acknowledgments

We would like to thank Fernando Galindo-Rueda for his expert comments and support of this research. This work was partially funded by the National Institutes of Health under awards P01AG039347, U01GM098959, and U01CA198934.

References

Booth, B., & Zemmel. R. (2004). Prospects for Productivity. *Nature Reviews Drug Discovery 3*, 451-456.

Jibu, M. & Osabe, Y. (2014). Refined R&D Indicators for Pharmaceutical Industry. *Future Information Technology, Lecture Notes in Electrical Engineering*, *309*, 549-554.

Lerner, J. (1994). The Importance of Patent Scope: An Empirical Analysis. *The RAND Journal of Economics*. 25(2), 319-333.

Lukach, R. & Plasmans. J. (2002). Measuring knowledge spillovers using patent citations: evidence from the Belgian firm's data. CESifo Working Paper NO.754 Category 9: Industrial organization,

OECD, (2013). *OECD Science, Technology and Industry Scoreboard 2013: Innovation for Growth.* Paris, France: OECD Publishing.

Squicciarni, M., Dernis, H. & Criscuolo, C. (2013). Measuring Patent Quality: Indicators of Technological and Economic Value. *OECD/DSTI/DOC* 3.

THEORY

METHODS AND TECHNIQUES

Can Numbers of Publications on a Specific Topic Observe the Research Trend of This Topic: A Case Study of the Biomarker HER-2?

Yuxian Liu[1,2,3], Michael Hopkins[2] and Yishan Wu[4]

yxliu@tongji.edu.cn, m.m.hopkins@sussex.ac.uk, wuyishan@istic.ac.cn

[1]Tongji University, Tongji University Library, Siping Street 1239, 200092 Shanghai (China);

[2]Tongji University Development & Planning Research Center, Siping Street 1239, 200092 (China);

[3]Univ. of Sussex, School of Business, Management and Economics, SPRU, Falmer BN1 9SL, Brighton (UK)

[4]Institute of Scientific and Technical Information of China, 15 Fuxinglu, Beijing 100038 (China)

Abstract

Using the accumulative publication data on HER-2 and its trend line, we draw the accumulative curve of the publication data. We discuss the characteristics of the accumulative publication curve, and how these characteristics change with respect to the different trend lines. We find that the points that regression line and the publication curve intersect with each other and the minimum points with respect to the trend lines do not change very much in both exponential trend line and linear trend line even if the exponential trend line raises itself much faster than the linear trend line. These data points are formed around the time when the significant discoveries are made and the related regulations are executed. These significant discoveries and regulations impact how and where the research should go and how the basic discoveries influence their application. The accumulative publication curve itself tells us very little about science. However the change of the accumulative publication curve with respect to the trend lines may tell us how science evolves. The content in the publications with significant scientific value may change the direction and trend of research, while research may change the publication trend the other way round. We may say that important scientific discoveries and regulations on clinical practice act as tipping points or act as drivers of change in the rates of scientific publications on the topic of HER-2. This induces us further to explore how scientific events drive the publication process. We may expect that through the publication process, we can monitor the scientific process.

Conference Topic

Theory

Introduction

The number of publications is widely used to measure the output or the productivity of researchers or their affiliated institutes. Hence, it is also used to compare the output of different countries (Bornmann & Marx, 2013; Zhu et al., 2004; Inglesi-Lotz & Pouris, 2011; Garfield, Pudovkin, & Paris, 2010). (China is ranked the second in terms of output of scientific research measured by the number of publications.) It is normally regarded as a quantitative indicator. The number of citations is supposed to measure the impact or the visibility of the researchers or their affiliated institutes that are investigated (Garfield, 1955). Sometimes it is even referred to as the indicator that measures the quality of the research in the cited article that a researcher has performed.

However, these measurements arouse a heated debate. In the December 16, 2012, the concerned scientists gathered in the Annual Meeting of the American Society for Cell Biology developed a set of recommendations referred to as the *San Francisco Declaration on Research Assessment* (DORA). DORA aimed to stop the use of the "journal impact factor" (JIF) in judging an individual scientist's work. They invited interested parties to indicate their support by adding their names to this declaration. Later the editor-in-chief of *Science* Bruce Alberts published an editorial to support this declaration. He thought the evaluation based on JIF was destructive and just encouraged "me-too science" and hence blocked innovation and created a strong disincentive to pursue risky and potentially groundbreaking work. Many leading scientists and scientific organization endorsed in this declaration (Alberts, 2013). JIF, a scientometric indicator based on the number of publications and the number of citations,

was originally created as a tool to help librarians to select journal to purchase, but later it is frequently used as a measure of the scientific quality of research in an article published in this journal and act as the primary parameter with which to compare the scientific output of individuals and institutions. Some academic institutes even use it to decide if a researcher should be funded or promoted as a tenure member (Garfield, 1999; Alberts, 2013). However, this practice arouses the fierce objection by scientists who are evaluated.

Bibliometricians also gave their voices to this phenomenon. Wouters, Glänzel, Gläser, & Rafols (2013) call for the urgent debate on the dilemmas of performance indicators of individual researchers. The Higher Education Funding Council for England (HEFCE), which distributes public money for higher education to universities and colleges in England and ensures that this money is used to deliver the greatest benefit to students and the wider public, carry out a work to review the role of metrics in the assessment and management of research. (http://www.hefce.ac.uk/whatwedo/rsrch/howfundr/metrics/). In the review, the working group launched a call for evidence to gather views and evidence relating to the use of metrics in research assessment and management. Elsevier and SPRU responded to the call. Ismael Rafols, Paul Wouters and Sarah de Rijcke organized a special session on the quality standards for evaluation indicators: Any chance for the dream to come true? (STI program). This session initiated to make the Leiden manifesto on the research assessment. van Raan, a scientometrics pioneer and gatekeeper (Garfield, Pudovkin, & Paris, 2010), will coordinate among different aspects so that this manifesto could be accepted widely. All these principles and responses, without exception, mention that quantitative information provided by metrics must be complemented by qualitative evidence to ensure the most complete and accurate input to answer a question. Even DORA recommended that the funding agencies should consider a broad range of impact measures including qualitative indicators of research impact, such as influence on policy and practice. DORA also recommended the publishers should make available a range of article-level metrics to encourage a shift toward assessment based on the scientific content of an article (DORA).

Garfield (1979, p. 62) illustrated that:

> "If the literature of science reflects the activities of science, a comprehensive, multidisciplinary citation index can provide an interesting view of these activities. This view can shed some useful light on both the structure of science and the process of scientific development."

However, can metrics drawn from publications and citations provide qualitative indicators that reveal the contents of the publications so that metrics can measure the way the contents of the publications influence policy and practice? Liu & Rousseau (2013, 2014) expounded that citation in essence is the interaction of the perspectives on a specific scientific phenomenon, hence can be used to reveal how the scientific phenomenon is understood. With the help of the regression line and a detrended curve, Liu & Rousseau (2012) show that the citation diffusion curve of an article containing a really original idea has an S-shape similar to the standard innovation diffusion curve. The convex part corresponds to the academic phase of the field that Kao's idea initiated, while the concave part corresponds to the technology dominated phase. The curve in the post-technology phase paralleled the regression line. The points of inflection correspond to the phase transition from academic to application research, while minima indicate a breakthrough in academic phase, and maxima indicate a breakthrough in the technology dominated phase. This implies that breakthroughs may directly influence the rate of change of the diffusion process while phase transfers may influence the rate of change implicitly. They claimed that the theory of diffusion process expounded in this article have the potential use of discerning breakthrough and turning points in an S & T area and finding social, technological, political and economic factors influencing the development of science. Can we use the number of publications on a specific topic to

observe the research trend? How the regression lines and the detrended forms of the publication curve tell us about the development of science? Can we discern the breakthrough and turning points between the academic phase and applied phase? Can we find social, technological, political and economic factors influencing the development of science? In this article, we will use the publications on Biomarker Her 2 to illustrate how the scientific activities on a specific research topic influence the publication process. With the help of the regression line and the detrended forms of the publication curve, we try to identify the breakthrough in this area and trajectory of translating research finding into diagnostic tools, medicines, procedures, policies and education. We will combine descriptive material on the development of the research domain with the publication growth - presents a model of interconnections of the publication and citation process, we analyze the cumulative publication curve and compare it to major events in the field. We will show that important scientific discoveries and regulation of clinical practice act as tipping points/ drivers of change in the rates of scientific publications on the topic of HER-2.

Data

After comprehensive literature research, we determined our search string:
TS=("CerbB2*" OR "CerbB-2*" OR "Cer-bB2*" OR "C-erbB2*" OR "Cer-bB-2*" OR "C-erbB-2*" OR "C-er-bB2*" OR "C-er-bB-2*" OR "Cerb B2*" OR "Cerb B 2*" OR "erbB2*" OR "erbB-2*" OR "er-bB2*" OR "er-bB-2*" OR "erb b2" OR "erb b 2" OR "HER2" OR "Epidermal growth factor receptor 2" OR "EGFR2" OR "CD340" OR "her 2")
These words include all the spelling variants related to the biomarker Human Epidermal Growth Factor 2. Among these words, “Her 2” is the only word that is not specific which may bring us some noising results because “her 2” can be used as in “her 2 children” which has nothing to do with Human Epidermal Growth Factor 2. Worse, children can be replaced by any nouns. Between her 2 and the nouns, any adjectives can be added in between. Even worse, since the Web of Science (WoS) ignores all punctuations, any punctuations can be added in between. Also one item that has “her 2 children” does not necessarily mean it is not what we need. Even the articles which deal with Human Epidermal Growth Factor 2 do not exclude the expression “her 2 children”. These situations make it very difficult for us to formulate an effective search string. However, we use the position information and its follow up to judge if these articles are related to the topic that we are searching by a program (Chavarro & Liu 2014, Lang, Liu & Chavarro, 2015), if it cannot be judged by a program, we judge it manually. We have got 98 articles that are not related to our topic. We downloaded all these data in 27 May 2014 and then excluded these 98 articles. Hence we get 30,056 articles. Since the gene of Her2/neu did not have a uniform name at the beginning when the scientists found this gene, we picked up some articles from the reference list of the early articles. And we exclude the articles published in 2014, and then we get 29,210 publications. Using these 29,210 records we do some bibliometric analysis.
The numbers of publications per year increase in roughly linearly. It is said that when a research topic turns to the application science, fewer and fewer publications will be published, instead, more and more patents will be approved. But in our case, it is the opposite, the research topic on HER-2 has already been applied in the diagnosis and therapy, the numbers of the publications on this topic do not decrease at all.

Table 1. Cumulative numbers of publications, the first and the second order differences.

Year	*cumulative numbers of publication*	*the first order difference*	*the second order difference*
1981	1	1	
1982	1	0	-1
1983	1	0	0
1984	3	2	2
1985	6	3	1
1986	12	6	3
1987	29	17	11
1988	68	39	22
1989	133	65	26
1990	261	128	63
1991	467	206	78
1992	763	296	90
1993	1126	363	67
1994	1581	455	92
1995	2046	465	10
1996	2530	484	19
1997	3048	518	34
1998	3624	576	58
1999	4312	688	112
2000	4996	684	-4
2001	5980	984	300
2002	7006	1026	42
2003	8141	1135	109
2004	9414	1273	138
2005	10922	1508	235
2006	12527	1605	97
2007	14196	1669	64
2008	16262	2066	397
2009	18633	2371	305
2010	21040	2407	36
2011	23500	2460	53
2012	26423	2923	463
2013	29210	2787	-136

Methodology: Regression Trend Lines and Detrended Curves of Time Series Data

Table 1 is a time series data. A time series is a sequence of data points, typically consisting of successive measurements made over a time interval. In informetrics, the time interval can be defined in different shift (Liu & Rousseau, 2008). Normally we make a scatter diagram to see whether data change linearly or nonlinearly. Then we make a regression analysis to find the best-fitting curve to see how the data change over time. We can get a regression equation to explain the

degree of association or the relationship between the data and time. Based on the equation that fits past data as well as possible, we can predict values of the variable at points other than the observation points.

The linear regression is the straight line. The curves of the nonlinear regression curves, depended on the regression equations, have different shapes. For example, the curve can be exponent curves if the regression equation is exponent function. The other possible curves can be logarithmic curve, power curve and multinomial curve. The straight line from the linear regression and the curve from the nonlinear regression are also called trend lines. Figure 1 show the exponential, multinomial, power, linear and logarithmic regression curves of data in Table 1.

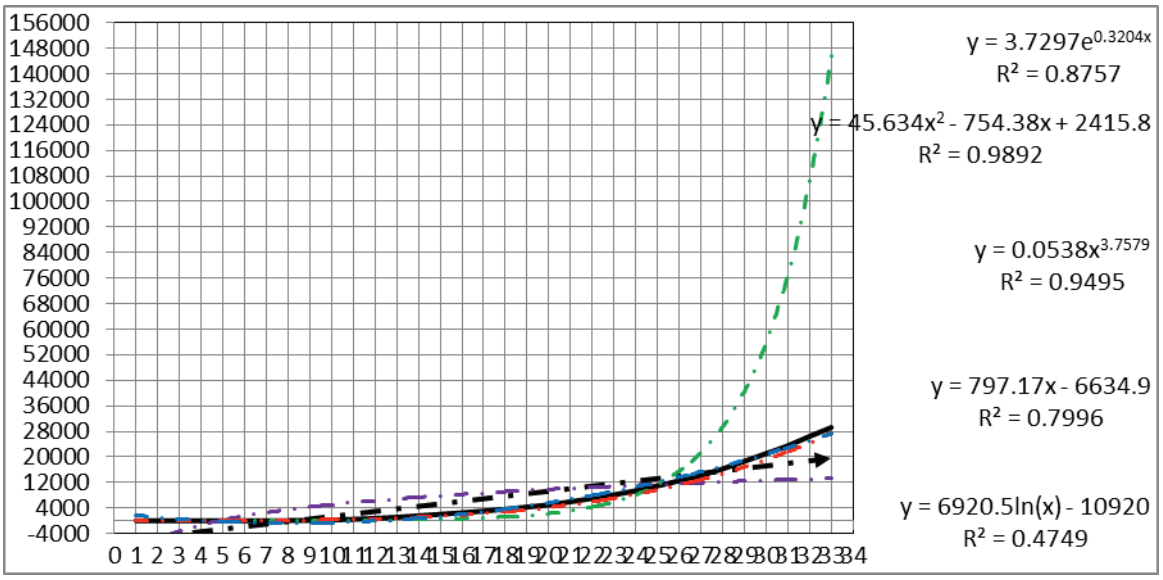


Figure 1.The exponential, multinomial, power, linear and logarithmic regression curves of data in Table 1.

Detrended curve (Shiavi, 1991) is a detrended description of the data. In order to draw a detrended curve, we find a trend line first and then calculating the difference between the overserved data and the trend line. It will give us a view on how the data change in terms of trend line. Peng et al. (1994, 1995) introduce the detrended fluctuation analysis. It is a scaling analysis method used to estimate long-range temporal correlations form. In other words, if a sequence of events has a non-random temporal structure with slowly decaying auto-correlations. It hence can eliminate the trend that self-affinity. By discerning long range correlation, it can help us understand what dominates the change of the data in the time series.

In this article, instead of calculating the difference between the observed data and the trend line, we will rotate abscissa to the paralleling line of the regression line and make the line touching the edge of the scatter diagram. The ordinate will pass through the first observation point so that all the numbers are positive. We then establish a new coordinate system. We will see how the data change with respect to the regression line

Results

We can choose different regression trend lines. In this article, we choose the best fitted straight line. Figure 2 shows the cumulative curve of the numbers of publications on her 2, its regression line and its minimum with respect to the regression line. We can see that the cumulative curve of the numbers of the publications on HER-2 is convex. The regression line intersects with the original data around 1987-1988 and 2007-2008. The minimum with respect to the regression line is around 1998-1999 (1 is the year 1981, 2 is 1982 and so on).

Now we know gene HER-2 was identified in 1981 by transfection studies with DNA from chemically induced rat neurogliobalstomas by Shih, Padhy, Murray and Weinberg (1981). From 1981 to 1987, several groups identified this gene independently (Schechter et al., 1985; Coussens et al., 1985; Semba, Kamata, Toyoshima, & Yamamoto, 1985; Fukushige et al., 1985). Slamon, Clark, Wong, Levin, Ullrich, and McGuire (1987) found correlation of relapse and survival with amplification of the HER-2 oncogene. HER-2 became a significant prognostic factor. Since then Slamon started to do research on binding to the HER-2 protein and prevents it from relaying a signal that stimulates the cancer cell to divide (Pioneers, 2007). In 1998, Herceptin was approved by FDA. Since then a revolutionary treatment started its journey in the history of human being to conquer the disease, based on the gene analysis, personalized treatment appear in the horizon that people can see. In 2007, American Society of Clinical Oncology (ASCO) and The College of American Pathologists (CAP) developed guidelines for when and how the status of HER-2 should be tested (Wolff et al., 2007). This guidelines were updated in 2013 (Wolff et al., 2013). Since then the test for the statues of HER-2 and clinical treatment with Herceptin become a standard test and treatment. However, as Herceptin did not take effect in some patients, the subpopulation remains to be defined, and side effects including cardiotoxicity need to be solved (Kumler, Tuxen, & Neilsen, 2014), HER-2 is still a topic that needs more investigations. We indicate these important events in Figure 2.

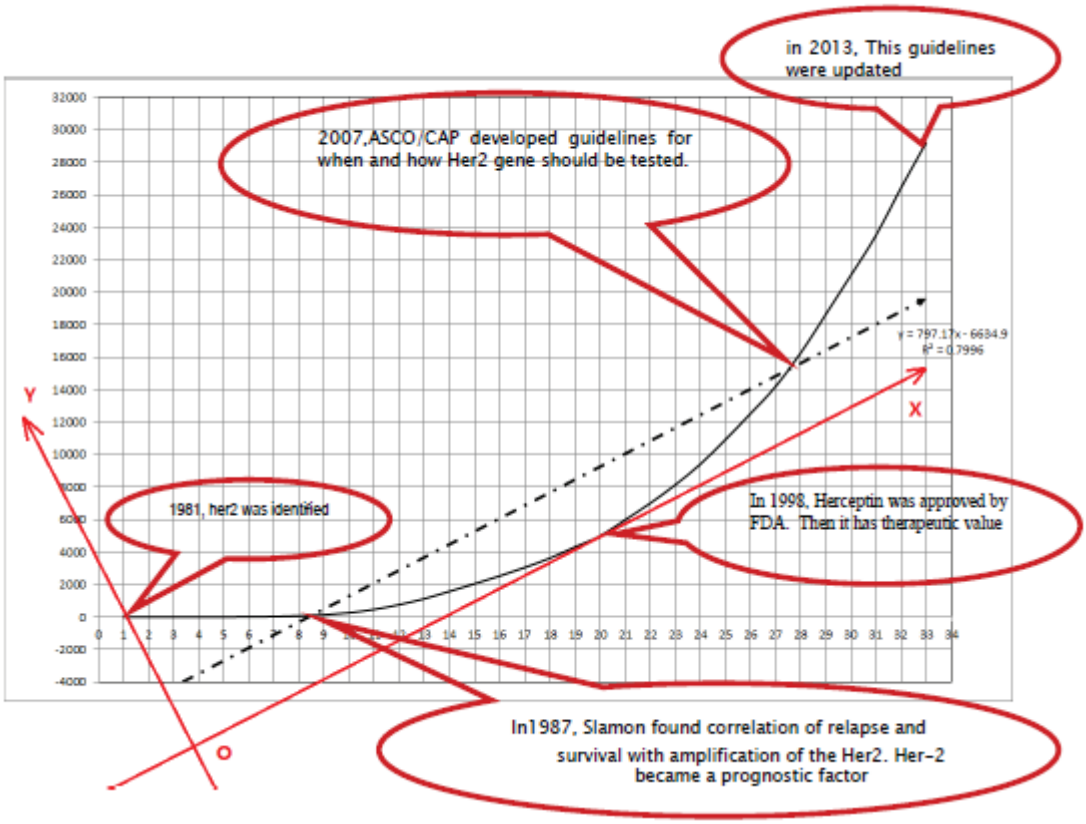


Figure 2. The cumulative curve of the numbers of publications on her 2, its regression line and its minimum with respect to the regression line.

We know that the minimum is a key point where the first-order derivative changes from negative to positive. If a curve shows the status of a thing that changes over time, we say it describes a kind of motion. The motion described by this curve changes from decreasing to increasing in the minimum point. A motion may have a different appearance as viewed from a different reference frame. If we choose the actual data as reference frame, to see how the trend changes, we can see that the discovery of the correlation of relapse and survival with amplification of her-2 oncogene in 1987 changes the trend reflected with publication data. This discovery made the amplification of her-2 a significant predictor and prognostic factor.

The numbers of publications start to increase significantly, the passion on this research topic is activated. Though with respect to the trend line, the original data curve decreases monotonically; the curve did not begin to increase until 1998 when Herceptin was approved by FDA. It is similar to the minimum point in cumulative number of citations curve of Kao when optical fiber was invented by Corning Glass Works in 1970. The crucial material problem, optical fiber, which Kao said in his conclusion "appears to be one, which is difficult but not impossible" was solved. This invention helped Kao realized his dream that no one believed it at the beginning (Liu & Rousseau, 2012). It is a coincidence that no one believed that the method Dr. Slamon used would work and the drug he created would be approved by FDA. On the contrary, everyone thought Dr. Slamon was crazy and he could not even find a student assistant majoring in science at the beginning (see the movie: Living proof and Bazell, 1998). In 2007, HER-2 test in breast cancer was recommended by ASCO-CAP, HER-2 research entered into another stage. The second order difference decreases after 2008. It dropped tremendously in 2010 and 2011. But in 2012, it went up tremendously which probably was caused by the fact that the recommendation guideline was challenged by the clinic practices and the new progresses. In 2013, ASCO/CAP convened an Update Committee that included coauthors of the 2007 guideline to conduct a systematic literature review and update recommendations for optimal HER-2 testing. In 2013, the second order difference become negative. Does the curve reach the point of inflection? We know the negative second order difference means the curve change from convex to concave. So far we cannot get to this conclusion. More observations are needed, at least we need to know how many publications on HER-2 will be published in 2014 so that we can judge whether it is an innate trend or just an occasional fluctuation. However, since major debate was settled down, though HER family oncogene (erbb1 erbb2, erbb3, erbb4) need to be dually blocked, and relative subpopulation needed to be defined and side effects refrain the use of some new developed medicine. For the moment there is an urgent need for prospective biomarker-driven trials to identify patients for whom dual targeting is cost effective (Kumler, Tuxen, & Neilsen 2014), we say it is not a major obstacle. We expect that the year when the breakthrough will make on these obstacles will appear in the maximum point on the curve drawn by the numbers of the publications on the HER-2. But it would depend on whether the research topic HER-2 gives rise to the other research topic.

The predictive, prognostic and therapeutic value of HER-2 are what changes the trend of research. The discoveries of these values of HER-2 influence the diffusion of the knowledge on HER-2 in the landscape of human intellectual space.

Selection of Trend Lines and the Different Implications that Detrended Line can Give Us

We can choose exponential, linear, logarithmic or power function as the trend lines to see what the data can tell us. Intuitively these trend lines are totally different, we hence imagine that the different trend lines can tell us totally different stories. But Figure 1 tells us the points that the different trend lines cross the data are slightly different, all around 2005-2008 even if the exponential trend is a much faster trend than the linear one. However, it is difficult to establish a new coordinate system to see clearly what the data tells us. Since the exponential curve is a straight line in semi-logarithmic system, we draw a scatter diagram in semi-logarithmic system (Figure 3). The data curve is concave upwards with respect to the exponential trend line. We can see the extremum with respect to the exponential trend line is around 1994, a little bit earlier than the time when the Herceptin was approved. However, it is in 1994 that Prof. Slamon finished phase 3 trial and was waiting for the decision of FDA. The first point that the trend line crossed the data is the same, but the second point is a little bit earlier. But the 2007 guideline was accepted for publication in September 27, 2006. The

expert panel was convened in 2005 and started to work on the guideline. It seems as if the shift of time is still in the acceptable region.

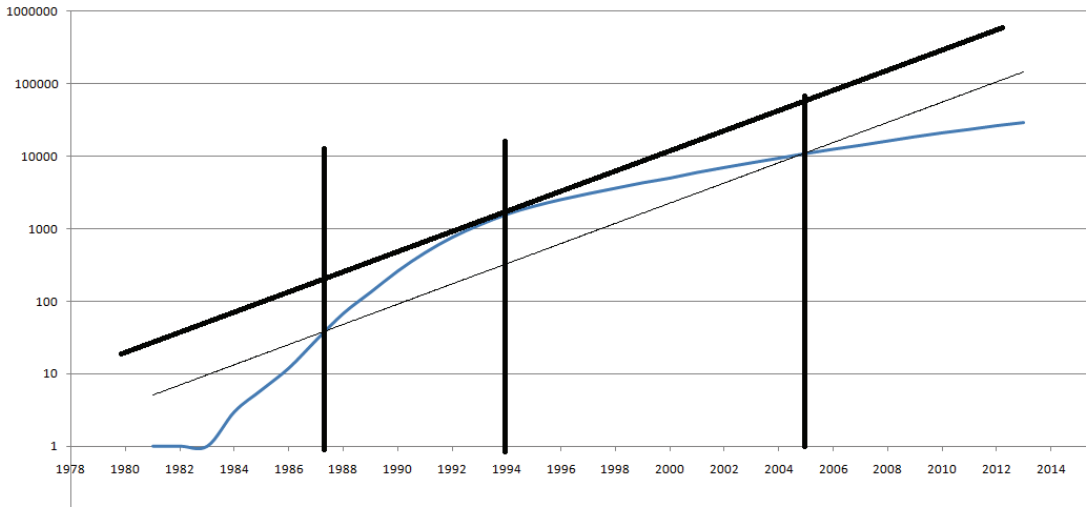


Figure 3. Cumulative publication data curve and its exponential trend line in the semi-logarithmic framework.

Publication and Citation Diffusion Process

Liu (2011) and Liu and Rousseau (2012, 2013, 2014) explore the determining factors that influence the citation process, and link the citation to the cognitive process of a scientific phenomenon under investigation. Through these articles, we illustrated the interaction of different perspectives on the phenomenon under investigation and how it is that the new ideas are accepted by academia determine the citation diffusion process.

In this investigation, we show that publication data curve with respect to the trend line can reflect how the important scientific events such as scientific discoveries and the release of government regulations in the clinical practice can change the trends of the publication process. Obviously, the primary knowledge creation process influences not only the citation process but also publication process. The change of research trends can show themselves in the publication data curve with respect to the trend line.

Liu and Rousseau (2010) studied two forms of diffusion, namely diffusion by publication and by citation. They tried to illustrate that publication diffusion is dominated by the internal diffusion mechanism that originates from the fact that a group of scientists expands their own (field) border. The citation diffusion is dominated by the external diffusion mechanism that the publication of the group of scientists, published in more and more fields, have potential to be applied in the other fields. Obviously, the publication diffusion process and citation diffusion process are interlinked with each other in that publication diffusion process determines the citation diffusion process.

As a matter of fact, publication process is entangled with citation process. Figure 4 shows how these two processes are entangled. Once the scientist(s) are interested in the scientific phenomenon, on the one hand they observe this phenomenon and get some preliminary impressions, and from these impressions they formulate some scientific ideas. On the other hand, they read the literature, which discusses this phenomenon and the perspectives to interact with the ideas that they formed by their observations to help them to get new insight into the phenomenon, and they then begin to make a thorough investigation. From these investigations scientists get new perspectives on the phenomenon. They articulate the new perspectives into a publication. When they write the manuscript they cite the old perspectives in the literature (perhaps they also read the other literatures for new evidence to convince the

readers). Publication and citation are thus born. In this process, scientific phenomenon is more and more clearly cognized.

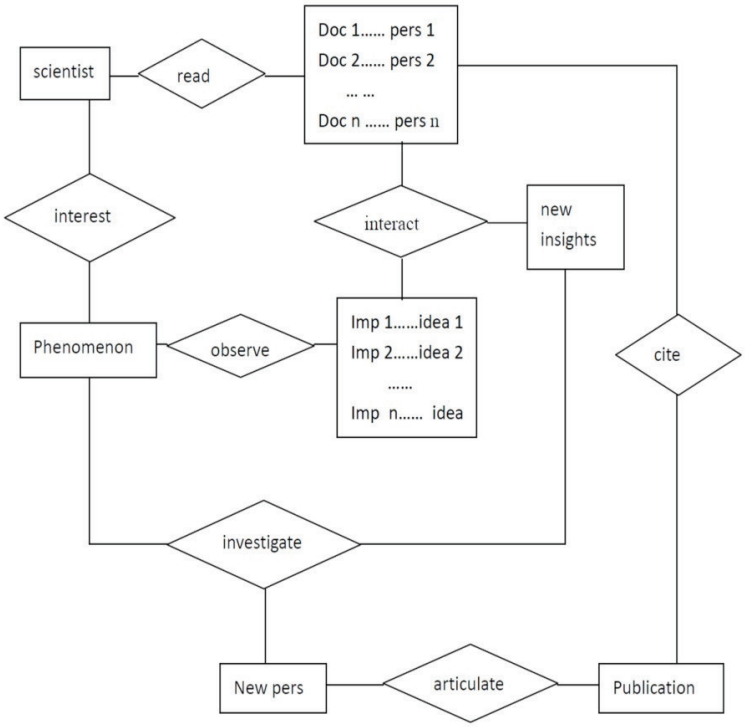


Figure 4. Entangled publication and citation diffusion process.

We can see from Figure 4 that the citation and publication processes are dynamic movement processes driven by the cognitive process of a phenomenon under investigation. The cognitive process is constituted (was led) by a series of scientific events. In this sense we may say that scientific events act as an engine to drive the evolution of science. Some events can lead the cognitive process to another direction. For example, the research in the publication Slamon, Clark, Wong, Levin, Ullrich and McGuire (1987) led HER-2 research from basic research to applied research. This event will change the research trend. Some events have no significant influence on the research trend.

Every scientific event could be represented by some publications on this event. In this sense, the scientific events drive the publication process, this process then drives the citation process. Scientific ideas that the publications convey are then diffused into the human intellectual landscape. So publication diffusion process may give us a deeper insight into the scientific events. The relationships between different publications are not as clear as that in citations, though through co-authorship or co-keywords we can establish different networks. But co-authorship or co-keywords did not reveal how the idea in one publication is diffused into the other publication. We cannot trace how the ideas in different publications interact with each other. Probably the mechanism of publication diffusion process needs to be explained via the citation diffusion process communicating different perspectives of the phenomenon under investigation. Therefore, citation and publication have a potential to reveal the cognitive process of the phenomenon under investigation.

However, we must understand how a scientific idea is diffused in the abstract intellectual landscape. This is the academic movement. In order to describe the academic movement we need to know where an idea comes from, where it will go, how fast the diffusion process is, how long is the distance from its start point to its destination. However, we face a lot challenges. First of all, we must mark the landscape with these scientific events. We have the

classification system such as the Library of Congress Classification System, Chinese Classification System, the WOS subject areas and the ESI fields. However, these systems alone cannot mark the scientific event. Because of the inaccuracy of this system, this kind of research does not give us more sense about the cognitive process of a research topic. Trochim and his colleagues (2011) proposed to identify "markers" in the translation process. They then assess the time that it takes for outputs to move across markers (Molas-Gallart, Este, Llopis & Rafols, 2014). Maybe this kind of mark system that embedded in a concrete scientific investigation will give us more information about the cognitive process of a scientific research.

Secondly, the distance in the human intellectual landscape may change over time and the destinations for the diffusion process are uncertain. These will make it very difficult to describe the scientific cognitive process via publication and citation diffusion process. These research questions deserve our effort. We would understand the scientific process more accurately if we could describe publication and citation diffusion processes more precisely. We can even anticipate what drives the evolution of science.

Conclusion

With the numbers of the publications on HER-2, we drew the accumulative curve of the publication data. We discuss the characteristics of the accumulative publication curve with respect to its trend lines and how its characteristics change in different trends. We find out the intersect points through regression line and the publication curve. These points are around the time when significant discoveries and regulations are made. These significant discoveries and regulations dominate how and where the research should go and how the basic discoveries influence their application. The accumulative publication curve itself tells us very little about how the science is evolving, but the change of the accumulative publication curve with respect to the trend lines may tell us more about the science. The content in the publication that has significant scientific value may change the direction and trend of research, hence change the publication trend reversely. We may say that important scientific discoveries and government regulations on clinical practice act as tipping points or act as drivers of change in the rates of scientific publications on the topic of HER-2. This makes us go further to explore how scientific events drive the publication process.

Acknowledgements

Yuxian Liu thanks Ismael Rafols, Raf Guns, Tim Engels, and Ronald Rousseau for the discussion in the early stage of this research. This work is supported by NSFC via 71173154. Yuxian Liu further acknowledges support from the China Scholarship Council.

References

Alberts, B. (2013). Impact Factor Distortions. *Science*, 340(6134), 787.

Bornmann, L. & Marx, W. (2013). Proposals of standards for the application of scientometrics in the evaluation of individual researchers working in the natural sciences. *Zeitschrift für Evaluation,* 12(1), 103-127.

Chavarro, D., & Liu, Y. (2014). How can a word be disambiguated in a set of documents: using recursive Lesk to select relevant records. http://www.gtmconference.org/pages/program.html

Coussens, L., Yang-Feng, T. L., Liao, Y. C., Chen, E., Gray, A., McGrath, J. et al. (1985). Tyrosine kinase receptor with extensive homology to EGF receptor shares chromosomal location with neu oncogene. *Science*, 230, 1132-1139.

Elsevier. (2014).Response to HEFCE's call for evidence: independent review of the role of metrics in research assessment. http://www.elsevier.com/__data/assets/pdf_file/0015/210813/Elsevier-response-HEFCE-review-role-of-metrics.pdf.

Fukushige, S. I., Matsubara K.I, Yoshida, M., Sasaki, M., Suzuki, T., Semba, K. et al. (1986). Localization of a novel v-erbB-related gene, c-erbB-2, on human chromosome 17 and its amplification in a gastric cancer cell line. *Molecular and Cellular Biology,* 6, 955-958.

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

Garfield, E (1979). *Citation Indexing.* New York: Wiley.

Garfield, E. (1955). Citation index for science. *Science*, 122, 108-111.

Garfield, E., Pudovkin, A. I., & Paris, S. W. (2010). A bibliometric and historiographic analysis of the work of Tony van Raan: a tribute to a scientometrics pioneer and gatekeeper. *Research Evaluation*, 19(3), 161-172.

Inglesi-Lotz, R., & Pouris, A. (2011). Scientometric impact assessment of a research policy instrument: the case of rating researchers on scientific outputs in South Africa. *Scientometrics*, 88(3), 747-760.

Kumler, I., Tuxen, M.K., & Neilsen, D. L. (2014). Anti-Tumour Treatment, A systematic review of dual targeting in Her-2 positive breast cancer. *Cancer Treatment Reviews*, 40, 259-270.

Lang, F., Liu, Y., & Chavarro, D. (2015). Improving accuracy in data collection: Can machine learning classification help? (in press)

Liu, Y.X. (2011). The diffusion of scientific ideas in time and indicators for the description of this process. Unpublished Doctoral Thesis, University of Antwerp, Belgium.

Liu, Y. X. & Rousseau, R. (2008). Definitions of time series in citation analysis with special attention to the h-index. *Journal of Informetrics,* 2(3), 202-210.

Liu, Y.X., & Rousseau, R. (2010). Knowledge diffusion through publications and citations: A case study using ESI-fields as unit of diffusion. *Journal of the American Society for Information Science and Technology*, 61(2), 340-351.

Liu, Y. & Rousseau, R. (2012). Towards a representation of diffusion and interaction of scientific ideas: the case of fiber optics communication. *Information Processing and Management,* 48(4), 791-801.

Liu, Y. & Rousseau, R. (2013). Interestingness and the essence of citation. *Journal of Documentation*. 69(4), 580-589.

Liu, Y. & Rousseau, R. (2014). Citation analysis and the development of science: a case study using articles by some Nobel Prize winners. *Journal of the American Society for Information Science and Technology,* 65(2), 281–289.

Peng, C-K., Buldyrev, S.V., Havlin, S., Simons, M., Stanley, H.E., & Goldberger, A.L. (1994). Mosaic organization of DNA nucleotides. *Physical Review E,* 49:1685-1689.

Peng, C-K., Havlin, S., Stanley, H.E., & Goldberger, A.L. (1995). Quantification of scaling exponents and crossover phenomena in nonstationary heartbeat time series. Chaos, 5:82-87.

Dr. Dennis Slamon's development of Herceptin revolutionized breast cancer treatment and accelerated research into therapies customized for each individual patient: the personal approach. (2007, Fall-Winter). *Triumph*, 13-16.

Schechter, A. L., Stern, D. F., Vaidyanathan, L., Decker, S. J., Drebin, J. A., Greene, M. I. et al. (1984). The neu oncogene: an erb-B-related gene encoding a 185,000-Mr tumour antigene. *Nature*, 312(5994), 513-516.

Semba, K., Kamata, N., Toyoshima, K., & Yamamoto, T. (1985). A v-erbB-related protooncogene, c-erbB-2, is distinct from the c-erbB-1/epidermal growth factor-receptor gene and is amplified in a human salivary gland adenocarcinoma. *PNAS,* 82 (19), 6497-6501.

Shiavi, R. (1991). *Introduction to Applied Statistical Signal Analysis*. Homewood, IL: Irwin, Aksen.

Shih, C., Padhy, L., Murray, M., & Weinberg, R. A. (1981). Transforming genes of carcinomas and neuroblastomas introduced into mouse fibroblasts. *Nature*, 290, 261-264.

Slamon, D. J., Clark, G. M., Wong, S. G. Levin, Ullrich, A., & McGuire, W. L. (1987). Human breast cancer: correlation of relapse and survival with amplification of the HER-2/neu oncogene. *Science*, 235, 177-182.

Truex, D., Cuellar, M., & Takeda, H. (2009). Assessing scholarly influence: using the Hirsch indices to reframe the discourse. *Journal of the Association for Information Systems,* 10 (7), 560-594.

Wolff, A.C., Hammond, M. E. H., Schwartz, J.N.,Hagerty, K.L., Allred, D.C., Cote, R. J. et al. (2007). American Society of Clinical Oncology/College ofAmerican Pathologists Guideline Recommendations for Human Epidermal Growth Factor Receptor 2Testing in Breast Cancer. *Arch Pathol Lab Med*, 131, 18-43.

Wolff, A.C., Hammond, M. E. H., Hicks, D. G., Dowsett, M., McShane, L.M. et al. (2013, November 1). Recommendations for Human Epidermal Growth Factor Receptor 2 Testing in Breast Cancer: American Society of Clinical Oncology/College of American Pathologists Clinical practice Guideline Update. *Journal of Clinical Oncology*, 31(31), 399-4013.

Wouters, P., Glänzel, W., Gläser, J. & Rafols, I. (2013). The dilemmas of performance indicators of individual researchers: an urgent debate in bibliometrics. *ISSI Newsletter*, 9(3), 48-53.

Zhu, X., Wu, Q., Zheng, Y. Z., & Ma, X. (2004). Highly cited research papers and the evaluation of a research university: A case study: Peking University 1974-2003. *Scientometrics*, 60(2). 237-247.

Founding Concepts and Foundational Work: Establishing the Framework for the Use of Acknowledgments as Indicators

Nadine Desrochers[1], Adèle Paul-Hus[1] and Jen Pecoskie [2]

[1] *nadine.desrochers@umontreal.ca, adele.paul-hus@umontreal.ca*
Université de Montréal, École de bibliothéconomie et des sciences de l'information, C.P. 6128, Succ. Centre-Ville, H3C 3J7 Montreal, QC (Canada)

[2] *jpecoskie@wayne.edu*
School of Library and Information Science, Wayne State University, Detroit, MI, 48202 (USA)

Abstract
Building on the concepts of the reward system of science and social capital, Blaise Cronin brought forth the idea that rewards in science are threefold, forming a triangle built from authorship, citations, and acknowledgements. Of these, acknowledgments are the hardest to grasp and evaluate. After nearly 45 years of multidisciplinary research on acknowledgments and a corpus of over 80 scientific contributions, there is still no consensus on the value of acknowledgments in scholarly communication. This study aims to further acknowledgments research with a meta-synthesis of the literature, establishing the theoretical framework for the use of acknowledgments as bibliometric indicators. Based on in-progress content analyses, broad categories emerge revealing contextual information crucial to the understanding of acknowledgments. Applying our framework on data from the Web of Science, further phases of this study will provide large-scale findings based on a multidisciplinary sample. From there, it will be possible to envision recommendations for the standardization and use of acknowledgments as indicators. However, grounding the study of acknowledgments in their underlying theoretical considerations and conceptual foundations will ensure these recommendations respect the diverse traditions of the scientific field.

Conference Topic
Theory

Introduction and background
It is a broadly recognized fact that the scientific field has a very "high degree of codification", to borrow the Bourdieusian phrase (Bourdieu, 1996, p. 226). How and when one is admitted into the academic community, how a researcher acquires credibility within the scientific realm, and what contributions turn a researcher into a renowned scholar are endlessly evaluated, measured, and scrutinized. This high degree of codification helps to both foster and assuage the paradox that underlies the use of empirical measures to define what remains an intrinsically nuanced and contextualized concept: scientific "success".

Merton (1973) presented the sociology of science with the reward system of science, its recognition paradigm, and the nepotistic undertones of the Matthew effect; Bourdieu reframed the concept of recognition to befit the concept of symbolic capital. Blaise Cronin brought forth the idea that these rewards are threefold, forming a triangle built from authorship, citations, and acknowledgements (Cronin, 1995; Cronin, 2005; Cronin & Weaver-Wozniak, 1993). These are all part of the *illusio*, which encompasses the stakes of the academic "game", its rules, and the very fact that its rewards are worth pursuing (Bourdieu, 1988, p. 56).

Of these rewards, acknowledgments are the hardest to grasp and evaluate; reasons range from lack of standardization to name-dropping and ambiguous wording (Cronin, 1995; Cronin, 2014), as well as the placement of acknowledgments, which can vary from in-text mentions to paratextual elements situated outside the body of the text (Genette, 1997). Researchers have also called for stricter policies to inform the use of acknowledgments, prescribe their form, offer conditions for inclusion, or establish their ethical ramifications (Brown, 2009; Chubin, 1975; Pontille, 2001). For example, while Cronin's research (Cronin, 1995) showed that in

most researchers' view, obtaining permission to thank is unnecessary, certain current editorial policies (e.g., PLOS ONE, 2015) require any acknowledging party to obtain the acknowledged party's permission. Extricating one aspect of acknowledgments is also not always straightforward. The "Funding Text" (FT) field of the Web of Science (WoS) database, indexed since 2008, is a telling example, since it often contains all things and people acknowledged, not just the agencies or institutions that provided funds to the project. That being said, the FT field of the WoS has opened new avenues for this research by making massive datasets available.

However, the literature heeds one important and overarching warning: after nearly 45 years of multidisciplinary study and a corpus of over 80 scientific contributions, there is still no consensus on the value of acknowledgments, no potential for meta-analysis within this corpus, and, despite common questions, no shared framework for further analysis, nor any clear recommendations for standardization. Given this situation, this study aims to further acknowledgments research with potential contributions to scientific policy guidelines (editorial and institutional) and research assessment (individual and disciplinary) in the scientometrics field, which has shown ongoing interest for acknowledgments as a potential indicator (Cronin & Weaver-Wozniak, 1992; Cronin, 2005; Díaz-Faes & Bordons, 2014).

In order to gain an understanding of where acknowledgments research had emanated from and where it is currently situated in the scientific ecology, an initial overview of the literature on acknowledgments was conducted, leading to the retrieval and document-level analysis of 115 scientific publications, which became the subject of a chapter submitted for inclusion in a book on theories in informetrics (Desrochers, Paul-Hus, & Larivière, in press).

This phase of the research established that the reward triangle can and should be studied, not only for its three constituting factors, but also for the relationships between them. It showed that the meeting point of citation and authorship is the apex of the reward triangle. Acknowledgements, however, are foundational in that they reveal the inner workings of the scientific *illusio* (Bourdieu, 1988) that support this apex and that have, historically, supported key conceptual frameworks: the "invisible college" (Crane, 1972), "trusted assessors," encountered before and during the peer review process (Mullins & Mullins, 1973), and the categorization of authors vs. acknowledged contributors (Patel, 1973).

Methodology

Following this initial review, it became clear that a meta-analysis of acknowledgments research would not be possible; however, the range of complex and varied approaches could form the basis for a meta-synthesis (Rousseau, Manning & Denyer, 2008) of the literature. This will: extract knowledge on the perceptions of acknowledgements across a variety of disciplines (e.g., Information Science, History, Astronomy, Literature, and Psychology); provide scientometricians with information pertaining to the nuances and contexts of research creation in various disciplines; and yield the conceptual framework necessary to undertake acknowledgements research on a larger scale using multidisciplinary datasets. The following research questions were thus devised:

1. What does "acknowledgment research" look like?
    a. Throughout history? (1970-present)
    b. What were its founding concepts and considerations?
    c. How are acknowledgments perceived and positioned in the acknowledgments literature itself?
2. Who is concerned with acknowledgment research?
    a. Scientists from what fields conduct acknowledgment research?
3. What aspects of acknowledgments are studied in acknowledgment research?

Using approaches based in the Social and Health Sciences (Rousseau et al., 2008; Dixon-Woods et al., 2005; Mays, Pope & Popay, 2005) and recommendations specific to the use of evidence-based literature in Information Science (Urquhart, 2010), a protocol for meta-synthesis was established using the PRISMA model for systematic literature reviews (Moher et al., 2009). The most recent searches place the corpus at 80 relevant documents. This paper presents preliminary findings and initial theoretical considerations.

Preliminary Findings and Discussion - Foundations for a theoretical framework

Based on in-progress content analyses, broad categories are emerging; they reveal contextual information crucial to the understanding of acknowledgments as potential bibliometric indicators.

Paratextual Status: Acknowledgements can be elusive, especially in structure-driven datasets. Standardized locations, conventions, separate paragraphs, in-text allusions, database fields defined as pertaining to one aspect but including others are all intrinsic to understanding their value.

Disciplinary Contexts: The literature stems from various disciplines, yielding a broad range of methods and reporting styles. It also approaches the topic from various angles: a discipline (e.g., Cronin, 2001), a culture or a group (e.g., Woolf, 1975), a linguistic community (e.g., Al-Ali, 2010), a specific journal or set of journals (e.g., Rattan, 2013), dissertations (e.g., Gesuato, 2004), or direct enquiry (e.g., Heffner, 1979), quantitative (e.g., Costas & van Leeuwen, 2012) or qualitative (e.g., Bashtomi, 2008). These differences do provide a spectrum of perspectives that need to be part of any standardization process of these scholarly rewards into contextualized indicators.

The Thankers and the Thanked: At its core, acknowledgments research is based on the basic questions of who or what gets thanked by whom and for what. From the expression of gratitude towards spouses to the mention of support from grant agencies, scientific acknowledgments reflect the same diversity as acknowledgments from other types of writers, such as literary writers (Desrochers & Pecoskie, 2014) and can be seen as a "'ledger' where debts are acknowledged" (Weber & Thomer, 2014, p. 84). Inconsistencies abound: people are thanked without specification of tasks, tasks are listed without names; financial capital is embedded with social capital and with messages of a highly personal nature (Coates, 1999).

Cloak and Dagger Reveals: The previous two categories show that scientific acknowledgments are sometimes as much a puzzle as they are clear; this in itself is information. Indeed, the last decades have shown interest in the fact that acknowledgments can expose the invisible college and pre-publication readers, including unknown reviewers, thereby setting boundaries between groups who know their identities and those who do not. This is obviously problematic in terms of using acknowledgments as indicators; yet abolishing this practice would mean revoking a practice that pays homage to the peer review process as it currently exists.

Language and Ethics: The acknowledgments genre has been studied in Linguistics and alluded to in other disciplines, including Information Science (Cronin, McKenzie & Stiffler, 1992). "How" entities are thanked is closely linked to prescribed funding-based requirements, cultural and disciplinary practices, and editorial guidelines, the latter being related to the ethics of thanks: securing permission to thank someone, paying 'lip service' to key players, and name-dropping (Cronin, 1995; Hollander, 2002)—angles reminiscent of the Matthew effect.

Value and Perception: Finally, acknowledgments research has the ingrained quality, seen elsewhere in science but perhaps rarely to this extent, to turn on itself. Numerous papers oscillate between two positions: perceiving acknowledgments as suitable for study and as potential indicators, true to the Merton-Bourdieu-Cronin theoretical continuum; and

criticizing them as problem-laden, lacking standardization, and fickle. Context and processes have come under scrutiny in the use of other indicators in research assessment; yet acknowledgment studies have a particular penchant for self-deprecation while relying on what is now four decades of research to insist upon the fact that there is something to this paratext.

Conclusion and Upcoming Phases

Quantitative content analysis will help weigh these concerns throughout the history of acknowledgments research. Qualitative analysis will help nuance these findings through context, history, and disciplinary boundaries. Together, these analyses will provide a meta-synthesis of the existing literature, from which the conceptual framework outlined here will be refined for use in further studies. The goal is to use this framework on data from the WoS and to provide large-scale findings based on a multidisciplinary sample. From there, it will be possible to envision recommendations for the standardization and use of acknowledgments as indicators.

However, since the literature provides many important warning signs, heeding them and grounding the study of acknowledgments in their underlying conceptual foundations will ensure these guidelines respect the multiple traditions of the scientific field and work within the boundaries of the evolving high stakes of codification. Furthermore, they will help take into account the fact that acknowledgments have long had a special standing in academia as the place where the *homo academicus* (Bourdieu, 1988) can make the invisible visible, but also vice-versa. This, in itself, is a stake of the *illusio* that deserves to be better understood.

Acknowledgements

This research was supported by the Social Sciences and Humanities Research Council of Canada. The researchers further thank Vincent Larivière for his support and insight.

References

Al-Ali, M. N. (2010). Generic patterns and socio-cultural resources in acknowledgements accompanying Arabic Ph.D. dissertations. *Pragmatics*, *20*(1), 1–26.

Basthomi, Y. (2008). Interlanguage discourse of thesis acknowledgements section: Examining the terms of address. *Philippine Journal of Linguistics*, *39*(1), 55–66.

Bourdieu, P. (1975). The specificity of the scientific field and the social conditions of the progress of reason. *Social Science Information*, 14(6), 19–47.

Bourdieu, P. (1988). *Homo academicus*. Stanford, Calif.: Stanford University Press.

Bourdieu, P. (1996). *The Rules of Art: Genesis and Structure of the Literary Field*. Stanford, Calif.: Stanford University Press.

Brown, R. (2009). How scholars credit editors in their acknowledgements. *Journal of Scholarly Publishing*, *40*(4), 384–398.

Chubin, D. E. (1975). Trusted assessorship in science: A relation in need of data. *Social Studies of Science*, *5*(3), 362–367.

Coates, C. (1999). Interpreting academic acknowledgements in English studies: Professors, their partners, and peers. *English Studies in Canada*, *25*(3-4), 253–276.

Costas, R., & van Leeuwen, T. (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.

Crane, D. (1972). *Invisible Colleges: Diffusion of Knowledge in Scientific Communities*. Chicago, IL: University of Chicago Press.

Cronin, B. (2014). Foreword: The penumbral world of the paratext. In N. Desrochers & D. Apollon (Eds.), *Examining Paratextual Theory and its Applications in Digital Culture* (pp. xv-xix). Hershey, PA: IGI Global.

Cronin, B. (1995). *The Scholar's Courtesy: The Role of Acknowledgement in the Primary Communication Process*. London: Taylor Graham.

Cronin, B. (2001). Acknowledgement trends in the research literature of information science. *Journal of Documentation*, *57*(3), 427–433.

Cronin, B. (2005). *The Hand of Science: Academic Writing and its Rewards*. Lanham, Maryland: Scarecrow Press.

Cronin, B., McKenzie, G., & Stiffler, M. (1992). Patterns of acknowledgement. *Journal of Documentation*, *48*(2), 107–122.

Cronin, B., & Weaver-Wozniak, S. (1992). An online acknowledgment index: Rationale and feasibility. In D. Raitt (Ed.), *Online Information 92: Proceedings of the 16th International Online Information Meeting, London, 5-10 December 1992* (pp. 281–290). Oxford: Learned Information.

Cronin, B., & Weaver-Wozniak, S. (1993). Online access to acknowledgements. *Proceedings of the 14th National Online Meeting 1993* (pp. 93–98). New York: M.E. Williams.

Desrochers, N., Paul-Hus, A., & Larivière, V. (in press.) The angle sum theory: Exploring the literature on acknowledgments in scholarly communication. In C. R. Sugimoto (Ed.), *Theories of Informetrics and Scholarly Communication.* Boston, MA: De Gruyter.

Desrochers, N., & Pecoskie, J. (2014). Inner circles and outer reaches: Local and global information-seeking habits of authors in acknowledgment paratext. *Information Research, 19*(1), paper 608. Retrieved from http://InformationR.net/ir/19-1/paper608.html

Díaz-Faes, A. A., & Bordons, M. (2014). Acknowledgments in scientific publications: Presence in Spanish science and text patterns across disciplines. *Journal of the Association for Information Science and Technology*, 65(9): 1834-1849.

Dixon-Woods, M., Agarwal, S., Jones, D., Young, B., & Sutton, A. (2005). Synthesising qualitative and quantitative evidence: A review of possible methods. *Journal of Health Services Research & Policy*, *10*(1), 45–53B.

Gesuato, S. (2004). Acknowledgments in PhD dissertations: The complexity of thanking. In C. Taylor Torsello, M. Grazia Bùsa, & S. Gesuato (Eds.), *Lingua inglese e mediazione linguistica. Ricerca e didattica con supporto telematico* (pp. 273–318). Padova: Unipress.

Heffner, A. G. (1979). Authorship recognition of subordinates in collaborative research. *Social Studies of Science*, *9*(3), 377–384.

Hollander, P. (2001). Acknowledgments: An academic ritual. *Academic Questions*, *15*(1), 63–76.

Mays, N., Pope, C., & Popay, J. (2005). Systematically reviewing qualitative and quantitative evidence to inform management and policy-making in the health field. *Journal of Health Services Research & Policy*, *10*(suppl 1), 6–20.

Merton, R. K. (1973). *The sociology of science: Theoretical and empirical investigations*. Chicago: University of Chicago Press.

Moher, D., Liberati, A., Tetzlaff, J., Altman, D. G., & The PRISMA Group. (2009). Preferred reporting items for systematic reviews and meta-analyses: The PRISMA statement. *PLoS Med*, *6*(7), e1000097.

Mullins, N. C., & Mullins, C. J. (1973). *Theories and Theory Groups in Contemporary American Sociology*. New York, NY: Harper and Row.

Patel, N. (1973). Collaboration in the professional growth of American sociology. *Social Science Information*, *12*(6), 77–92.

PLOS ONE. (2015). PLOS ONE manuscript guidelines: Acknowledgments. Retrieved from http://www.plosone.org/static/guidelines#acks

Pontille, D. (2001). L'auteur scientifique en question: Pratiques en psychologie et en sciences biomédicales. *Social Science Information*, *40*(3), 433–453.

Rattan, G. K. M. (2013). Acknowledgement patterns in annals of library and information studies 1999-2012. *Library Philosophy and Practice*, *e-journal* (paper 989). Retrieved from http://digitalcommons.unl.edu/libphilprac/989/

Rousseau, D. M., Manning, J., & Denyer, D. (2008). *Evidence in management and organizational science: Assembling the field's full weight of scientific knowledge through syntheses* (SSRN scholarly paper 1309606). Rochester, NY: Social Science Research Network. Retrieved from http://papers.ssrn.com/abstract=1309606

Urquhart, C. (2010). Systematic reviewing, meta-analysis and meta-synthesis for evidence-based library and information science. *Information Research, 15*(3), paper 708. Retrieved from http://www.informationr.net/ir/15-3/colis7/colis708.html

Weber, N. M. & Thomer, A. K. (2014). Paratexts and documentary practices: Text mining authorship and acknowledgment from a bioinformatics corpus. In N. Desrochers & D. Apollon (Eds.), *Examining Paratextual Theory and its Applications in Digital Culture* (pp. 84-109). Hershey, PA: IGI Global.

Woolf, P. (1975). The second messenger: Informal communication in cyclic AMP research. *Minerva*, *13*(3), 349–373.

Analysis on the Age Distribution of Scientific Elites' Productivity: A study on Academicians of the Chinese Academy of Science

Liu Jun-wan[1] and Zheng Xiao-min[2] and Feng Xiu-zhen[3] and Wang Fei-fei[4]

[1]*liujunwan@bjut.edu.cn,* [2]*xiaominzheng2014@sina.com,* [3]*xfeng@bjut.edu.cn,* [4]*feifeiwang@bjut.edu.cn*
School of Economics & Management, Beijing University of Technology, Beijing 100124, (China)

Introduction

Is there any regularity in scientists' research activities? For example, does there exist a period when a scientist makes his most contributions? If so, which period is the most productive period? To answer the questions above, many scholars have been contributed their efforts on studying the relationships between productivity and age, such as: (1) age distribution of scientists' creativity or productivity (Liming et al., 1996; Bonacarsi & Daraio, 2003; Jones 2010); (2) the relationship between the longevity and scientist's outputs (Levin & Stephan, 1991; Jonesa & Weinberg 2011; Todorovsky, 2014); (3) the effects of age on researcher's productivity (Bonacarsi & Daraio Costas & van Leeuwen, 2010). However, the previous research still leave some gaps need to be filled. One of them is what about the age distribution of an individual researcher's achievements in his research career. Our research efforts in this paper would contribute to this topic. Particularly, the object of our study is Academicians of the Chinese Academy of Science. And we explore the age distribution of publication by these academicians.

Data and Method

The website of Academic Divisions of the Chinese Academy of Sciences provides academicians' brief introduction and research experience, which including their birth day and affiliated institutions. We choose total 139 Academicians in field of *Mathematics & Physics*, and total 85 Academicians in field of *Information Technical Science* as our research data. *Mathematics & Physics* is an ancient and classical subject, and *Information Technical Science* is a rapid development subject. In order to analyze the age distribution of these academicians' publication, the academician's name and affiliation were used as joined retrieval terms to get their publications both in China National Knowledge Infrastructure (CNKI) and web of science (SCI) database. CNKI is the largest authoritative digital publishing platform and knowledge services platform in China. To get their whole publication output, the repetitive or mistaken publication data of these academicians were deleted.

The average age of 224 academicians is 74 years old, and all of these academicians are now alive until the retrieval day (11/2014). The number of the scientists' publications was selected as the scientific productivity indicator, but the co-author situation was equally considered. This paper considers age distribution of scientists' publication from the scientists' physiological age view.

Age distribution of academicians' publication

Firstly, we count the number of every individual academician's publication according to his physiological age. After that we sum the number of publication up according to the same physiological age of all academicians in the same field. So we can get the physiological age distribution of publication of total scientists in one field. We named papers indexed in SCI/CNKI as "SCI/CNKI" paper for short.

Age distribution of academicians' publication in Mathematics & Physics

The publication age distribution curve of CNKI paper and SCI paper of academicians in *Mathematics & Physics* are shown in Figure 1(a). The publication age distribution curve of total paper (sum of number of CNKI paper and SCI paper) is presented in Figure 1(b). Just as shown from the folder part of the two publication age distribution curves in Figure 1(a), we can see the period between the age of 50 and 65 is the same publication peak period of CNKI paper and SCI paper. Scientists published 61% of their total publications between the age of 50 and 71, the highest peak point is at the age of 68.

Age distribution of academicians' publication in Information Technical Sciences

The publication age distribution curve of CNKI paper and SCI paper of academicians in *Information Technical Sciences* are presented in Figure 2(a). The age period from 60 to 70 is the same publication peak period of CNKI paper and SCI paper. As Figure 2(b) is shown, scientists published 51% of their total publications between the age of 62 and 76, and the highest peak point is at the age of 67. In detail, there is a smaller publication peak period between the age of 45 to 51 before the higher one.

Significant differences test of academicians' productivity before and after tenure

Paired-Samples T Test was used to test if the scientists' productivity would be different before and after tenure. We sum up the number of publications for five years of every individual academician before and after tenure. Before testing, we assume that there is no significant difference of academicians' productivity before and after tenure, then we use the Paired-Samples T test to test the hypothesis. According to the analysis results, the assumption is rejected, which means that the number of publication is obviously different before and after tenure. After tenure, academicians are more productive than before in overall.

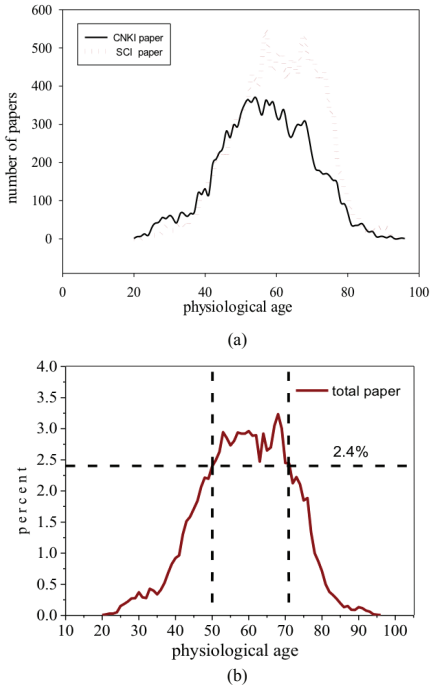


Figure 1. Publication age distribution of academicians in *Mathematics & Physics*.

Discussion and conclusion

The final results show that age distributions of academicians' publication have some regular features. The entire publication age curve of *Mathematics & Physics* shows a single peak distribution. The publication peak period is between the age of 50 and 71. However, publication peak period of academicians in *Information Technical Sciences* is between the age of 62 and 76. Moreover, it is different from *Mathematics& Physics,* which has a small publication peak period between 45 and 51 in publication age curve of *Information Technical Sciences'* academicians. Additionally, our results also reveal that there is significant difference of the scientists' productivity before and after tenure. The publication age distribution law on academicians of the Chinese Academy of Science brings us useful enlightenment. We should pay more attention to middle-aged scientists to improve their research input-output ratio.

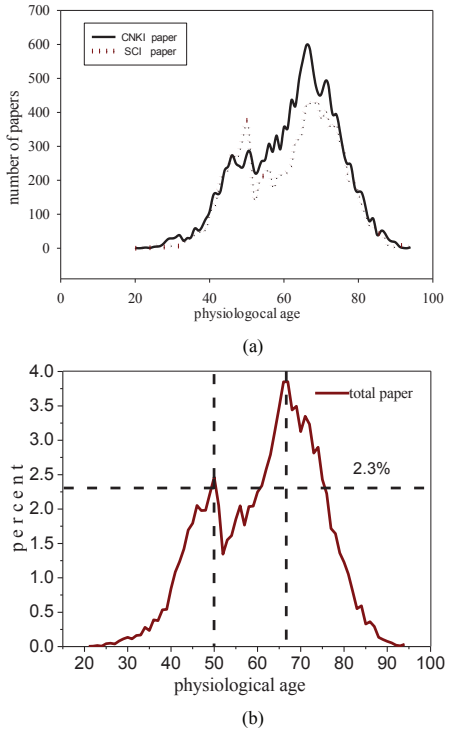


Figure 2. Publication age distribution of academicians in *Information Technical Sciences*.

References

Bonaccorsi, A., & Daraio, C. (2003). Age effects in scientific productivity. *Scientometrics*, *58*(1), 49-90.

Costas, R. & Van Leeuwen, T. N. (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. *JASIST*, *61*, 1564–1581.

Jones, B.F. (2010). Age and great invention. *Rev Econ Stat*, *92*, 1–14.

Jonesa, B. F. & Weinberg, B. A. (2011). Age dynamics in scientific creativity. *Proceedings of the national academy of sciences of the united states of America*, *108*, 18910-18914.

Levin, S.G. & Stephan, P. E. (1991). Research productivity over the life cycle: evidence for academic scientists. *American Economic Review*, *81*, 114–132.

Liming, L., Hongzhou, Z. & Yuan, W. (1996). Distribution of major scientific and technological achievements in terms of age group — Weibull distribution. *Scientometrics*, *36*, 3-18.

Todorovsky, D. (2014). Follow-up study: on the working time budget of a university teacher-45 years self-observation. *Scientometrics*, *3*, 2063-2070.

An Experimental Study on the Dynamic Evolution of Core Documents

Lin Zhang[1], Wolfgang Glänzel[2], Fred Y. Ye[3]

[1] *zhanglin_1117@126.com*
[1]Dept. Management and Economics, North China University of Water Conservancy and Electric Power, Zhengzhou (China)

[2] *Wolfgang.Glanzel@kuleuven.be*
[2]Centre for R&D Monitoring (ECOOM) and Dept. MSI, KU Leuven, Leuven (Belgium)
Dept. Science Policy & Scientometrics, Library of the Hungarian Academy of Sciences, Budapest (Hungary)

[3] *yye@nju.edu.cn*
[3]School of Information Management, Nanjing University, Nanjing 210023, (China)
Jiangsu Key Laboratory of Data Engineering and Knowledge Service, Nanjing 210023 (China)

Introduction

The concept of the core of documents had originally been introduced in connection of co-citation analysis (Small 1973). The term *core documents* has later been re-introduced in the context of bibliographic coupling (BC; see Glänzel & Czerwon, 1996) and hybrid BC and text based similarities (Glänzel & Thijs, 2011) in order to identify strongly interlinked papers that form important nodes in the network of scholarly communication. In order to study stability and dynamics of core-document sets we apply two different methods to h-index related literature in the period 2005–2013 for illustration.

Data Sources and Processing

Data were retrieved from Thomson Reuters Web of Science Core Collection (WoS) following the strategy of Zhang et al. (2011), with extension of the period 2005–2013. We also added citing papers but removed duplicates and papers with less than 5 references to avoid biases in BC similarities. We obtained a final set of 3,270 documents. Figure 1 shows the annual increment of papers in this set.

Research Questions, Methods and Results

In this study we apply two different methods to determine core documents, (Method I) the traditional one according to Glänzel & Czerwon (1996) with a fixed number of links ($n = 15$) and Method II using the h-core of the network (Glänzel, 2012). In both cases we applied a *hybrid approach*. We used link strengths of 0.5 and 0.4 according to Salton's cosine measure. Using these parameters, we analysed the dynamics of core documents along the following questions.

- How is evolution of core documents reflected by the two methods?
- Do the two methods provide stable results?
- Do core documents adequately represent the evolution of the topic?

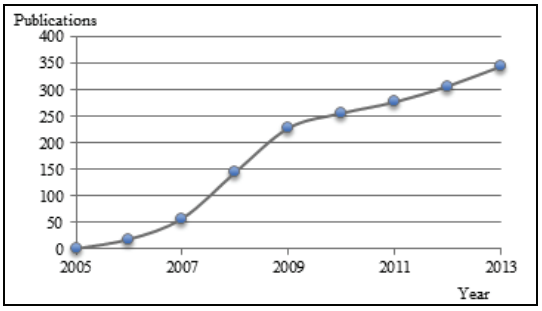


Figure 1. Distribution of h-related publications during 2005-2013.

Core document are by definition strongly interlinked with a large number of other documents in the set under study and thus represent the very core of the set. As expected, their number increases with expanding time spans, the average annual growth rate of the cumulative set amounted to 46% (Method I) and 25% (Method II), respectively. Not only the number of nodes in the network but also the number of their links is growing, however at a different pace. Indeed, we found that the complete h-related set increased at a large constant pace of 11% while the growth of the core sets was faster (see above), but its growth slowed down. This might in part be a consequence of the increasing age of references. In 2013 the core reached a representation of 2.0% and 2.4%, respectively. This characterizes the evolution of the core set with respect to the topic dynamics. The second question that arises from these figures is in how far do both methods mirror the same "core" of literature. In order to check the robustness of these methods, we compared the overlap of the sets of core documents obtained from the two methods. To this end we used BC with fixed number of links as reference standard. Concordance with Method I ranged between 83.8% and 95.2% with increasing trend from 2005–2007 to 2005–2013 and using Method II the shares ranged between 96.8% and 80.7%, however with decreasing trend.

In order to answer the third question, we analysed the core sets obtained from the two methods on the basis of authors and topics of the individual papers. The evolution of the core-document sets according to Method II is shown at three different stages in Figure 2 using Pajek with Kamada–Kawai layout (Batagelj & Mrvar 2003).

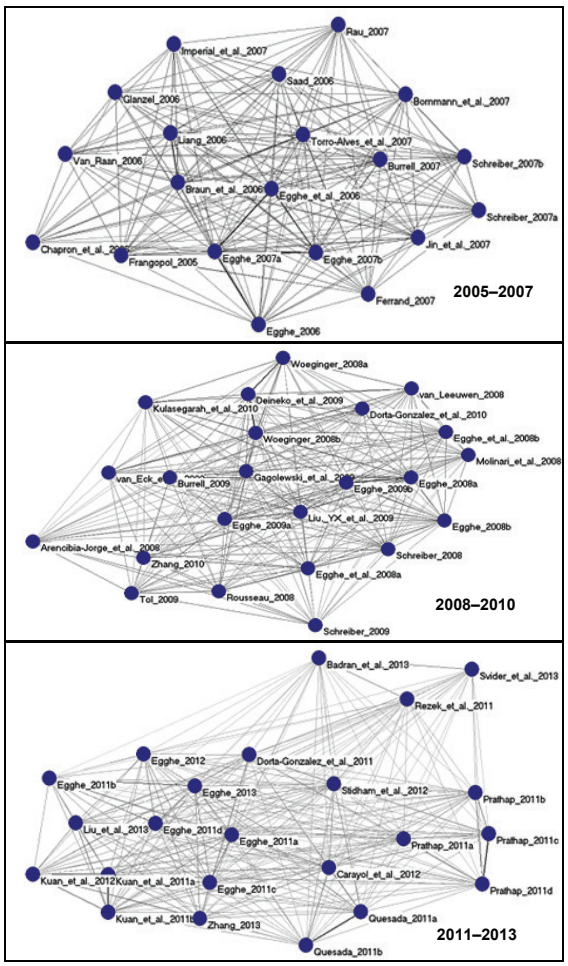


Figure 2. The evolution of the core-documents set (II).

Core nodes in Figure 2 are based on BC but hybrid similarities are used to measure the links between the nodes. This can be done because of the strong concordance between the sets obtained from the two methods. The links between core nodes in Figure 2 are denser and stronger than in the BC approach, which is due to the inclusion of textural information. The interpretation of Figure 2 is not straightforward, but the structural changes of the networks during different periods presented here are quite clear and noteworthy. The network in the first sub-period (2005–2007) comprises above all theoretical publications. The network of 2008–2010 already reflects a different picture. While most theoretical papers are still located in the centre of the network, also 'applied studies' started to appear in the core-documents set. These are distributed at the periphery of the network, which indicates that the topic starts to expand from pure theory to more application. The network of the last sub-period (2011–2013) reflects the clearest structure, where we could distinguish several sub-networks. As the most stable contributor, Egghe's six papers are found in one strongly interlinked sub-network, with the most theoretical roots. Unlike the network in 2008–2010, where some 'applied studies' were still scattered at the periphery of the network, we found more distinct sub-networks on 'applied' research in the network in 2011–2013. In this sense, core documents appear to follow the trend of the topic that is moving away from 'hard-core' informetrics towards research evaluation at different levels of aggregation and for various purposes.

Discussion and Conclusions

In the present study we focussed on 'core documents' with their evolution in publication networks using the example of a specific but nonetheless heterogeneous paper set. The two applied methods proved robust and representative. Their coverage amounted to about 2% of the topic literature, which is in line with the expectations (cf. Glänzel, 2012) but their links lead to related documents that represent a much broader coverage of the topic h-related literature.

The evolution of the core-document network represents the general tendency of shifts in topic, authors and application in an adequate manner. This gives also evidence that Hirsch-type indices have become a tool that is used also outside the informetric community.

Acknowledgments

Lin Zhang acknowledges the NFSC Grant No 71103064, Fred Ye acknowledges the NFSC Grant No 71173187 and Jiangsu Key Laboratory Fund for financial support.

References

Batagelj, V., & Mrvar, A. (2003). *Pajek–Analysis and visualization of large networks*. In M. Jünger & P. Mutzel (Eds.), Graph drawing software (pp. 77-103). Berlin: Springer.

Glänzel, W., & Czerwon, H. J. (1996). A new methodological approach to bibliographic coupling and its application to the national, regional and institutional level. *Scientometrics*, 37(2), 195–221.

Glänzel, W. & Bart Thijs, B. (2011). Using 'core documents' for the representation of clusters and topics. *Scientometrics*, 88(1), 297-309.

Glänzel, W. (2012). The role of core documents in bibliometric network analysis and their relation with h-type indices. *Scientometrics*, 93(1), 113–123.

Small, H. (1973). Cocitation in scientific literature – new measure of relationship between 2 documents. *JASIS*, 24(4), 265-269.

Zhang, L., Bart Thijs, B. & Glänzel, W. (2011). The diffusion of H-related literature. *Journal of Informetrics*, 5(3), 583-593.

How Related is Author Topical Similarity to Other Author Relatedness Measures?

Kun Lu[1], Yuehua Zhao[2], Isola Ajiferuke[3] and Dietmar Wolfram[2]

[1] *kunlu@ou.edu*
School of Library and Information Studies, University of Oklahoma, 401 West Brooks, Norman, OK 73019 (United States)

[2] *{yuehua,* dwolfram}*@uwm.edu*
School of Information Studies, University of Wisconsin-Milwaukee, P.O. Box 413, Milwaukee, WI 53201 (United States)

[3] *iajiferu@uwo.ca*
Faculty of Information and Media Studies, University of Western Ontario, London, ON N6A 5B7 (Canada)

Abstract
Using a dataset of 26,228 Psychology document surrogates from Elsevier databases, we compare author relatedness measure outcomes for 125 authors based on topic modelling to more traditional approaches that rely on direct citation, co-citation and collaboration. Outcomes for the author topical similarity measure are compared to existing co-authorships in the dataset using UCINET/NetDraw. We demonstrate how author topical similarity outcomes provide a similar, but more complete, picture of author relationships than the co-authorship network. Nonparametric correlation analysis results of author topical similarity, co-authorship, citation, and co-citation were also compared for thirty author pairs of differing author topical similarity values. There is a significant correlation between author topical similarity and co-authorship and direct citation-based measures for high similarity author pairs, but not with co-citation measures. The author topical similarity measure, therefore, may serve as a reasonable predictor of collaboration or direct citation for authors with high topical similarity. The measure may also identify potential collaborators based on high author pair similarity values, where there is a lack of existing collaboration, and serve as a complement to author relatedness based on co-citation analysis.

Conference Topic
Methods and techniques

Introduction
Understanding the relationships between authors is of great interest to researchers in scholarly communication and informetrics. Author relatedness can be revealing of the membership of research communities and potentially hidden similarities among authors that may not be readily apparent. The relatedness of authors is a multi-faceted concept that can be determined from different data sources, which include direct author citations, author co-citations (White & McCain, 1998), author bibliographic coupling (Zhao & Strotmann, 2014), author topical similarities (Lu & Wolfram, 2012), author collaborations (Glänzel & Schubert, 2005), and other derived measures (Jacobs & Wolfram, 2014; Jeong, Song, & Ding, 2014). These measures can be discursively categorized into three groups: citation-based (author citation or co-citation), content-based (author topical similarities), and collaboration-based measures (co-authorship). Among developed measures, citation-based measures, especially based on author co-citation analysis, are most influential and well-studied in the literature. The emergence of topic modelling techniques (Rosen-Zvi et al., 2010) has reheated the interest in content-based measures. Co-authorship has been widely used to understand scientific collaborations and reveal research communities. It is well understood that these measures focus on different aspects of author relationships and reveal different types of relatedness. However, the inter-relationships among the different measures have been rarely researched. Are authors with higher topical similarities more likely to collaborate with each other? Do they tend to cite

each other more often? Are they more likely to be co-cited by others? These questions are not adequately addressed in the literature. The purpose of this study is to examine the inter-relationships among several measures of author relationships, including citation-based measures, content-based measures, and collaboration-based measures. More specifically, the research aims to address the following questions:

1) Does author relatedness assessed by author topical similarity reveal similar relationships as a more traditional assessment approach based on co-authorship?
2) What is the relationship between author citation, author co-citation, author collaboration and author topical similarity?
3) Can author topical similarity be used as a predictor for other relatedness measures such as author collaboration, author direct citation or author co-citation?

Author topic modelling (Rosen-Zvi et al., 2010) will be used to determine author topical similarity using bibliographic records for the field of Psychology. Understanding the inter-relationships among the different author relatedness measures contributes to the better use of them in revealing scientific structures.

Literature Review

The present study builds on existing research examining author similarity comparison by employing topic modelling techniques and comparing outcomes to citation and co-citation-based measures.

Measuring the relatedness between scientific entities (e.g. articles, authors, and journals) has been studied for years. Typically, most similarity measures between units are based on quantifiable assessments arising from citation practices that link authors or through direct collaboration or other co-occurrence similarities (Börner, Chen, & Boyack, 2003). To date, the relatedness or similarity between authors has been investigated mainly through five perspectives: direct citation, bibliographic coupling analysis, co-citation analysis, co-authorship analysis, and co-word analysis. Direct citation relationships are built on citation behaviour when one author cites others' work (Boyack & Klavans, 2010). Bibliographic coupling relationships are measured by counting the same references two authors share in their publications and have been studied recently by Zhao and Strotmann (2008, 2014). Moreover, the most widely studied approach, co-citation analysis, assesses the association between two authors by the frequencies they were co-cited by others (White & McCain, 1998). A co-authorship relationship results from a direct collaboration (Glänzel & Schubert, 2005). Each of these methods relies on an explicit connection arising from citation or collaboration. Without these connections, no relationship can be identified. Implicit relationships can be revealed by comparing the content of documents authors have published. Until recently, this has taken the form of co-word analysis, where words or index terms from documents are used to determine how closely related entities of interest are (e.g., Law & Whittaker, 1992).

Although previous studies have used content-based methods to approach the relationships between authors, documents and disciplinary areas, topic-based methods have rarely been applied to date to capture the relationships between authors (Lu & Wolfram, 2012). Topic modelling seeks to automatically reveal the latent topics from a set of documents through machine learning. Hofmann (1999) first proposed a generative data model—called the Probabilistic Latent Semantic Indexing (PLSI)—that represented each document as a probability distribution over a set of topics. While Hofmann's work provided some advantages for document indexing, it may lead to serious problems of overfitting (Blei, Ng, & Jordan, 2003). To overcome the limitations of PLSI, Blei, Ng, and Jordan (2003) presented a three-level hierarchical Bayesian model, which is known as Latent Dirichlet Allocation (LDA). In the LDA model, each document is modelled as a finite mixture over an underlying

set of topics, where each topic is modelled as a mixture over an underlying set of terms (Blei et al., 2003). Follow-up efforts to extend content-level LDA modelling have been investigated using different approaches, such as the Author-Conference-Topic (ACT) model (Tang et al., 2008), correlated topic model (CTM) (Blei & Lafferty, 2006), interactive topic modelling (Hu, Boyd-Graber, Satinoff, & Smith, 2014), and supervised Latent Dirichlet Allocation (sLDA) (Mcauliffe & Blei, 2008). Most topic modelling studies explored the relationships between documents and topics. However, few studies have employed topic modelling methods to conduct author similarity comparison. The present study explores how topic modelling-based author relatedness assessment may complement existing methods based on citation and collaboration-based measures.

Method

Data collection

Elsevier, Inc. has provided a dataset consisting of selected data for 56,620 bibliographic records from 118 Elsevier Arts & Humanities journals. Initially, the authors explored the use of all the data, representing many disciplines within the humanities and social sciences. Outcomes using the author topic modelling approach outlined below resulted in inclusive topical assignments, likely due to the broad vocabulary represented that resulted in topical assignments that combined terms from different disciplines. The subset of the data assigned with Scopus subject classification code 3200, corresponding to "Psychology (all)", was used in this study. The Psychology subset represented the most frequent field appearing in the dataset.

The Psychology subset includes bibliographic records of 26,228 publications written by 63,695 different authors. The authors were identified using the *author_id* field included with the data. An Author-Topic LDA model (Rosen-Zvi et al., 2010) was trained on the title and abstract fields of the psychology subset. The number of topics (k) was set to 100 for exploratory purposes. Other parameters of the model were set as follows: alpha equals 0.5 ($50/k$), beta equals 0.01 and the number of iterations is 1000. All terms were normalized to lower case before processing. A standard list of English stop words were removed and Porter stemming was applied when processing the text. The descriptive statistics of the psychology subset are provided in Table 1. The document length is measured by the number of word tokens in the title and abstract after removing stop words (i.e. common words that were excluded). During the process, we found some authors were listed multiple times in an article because of their multiple affiliations. This was counted as one occurrence in the study.

Table 1. Descriptive statistics of the psychology subset (title and abstract fields).

Measure	*Frequency/Value*
# of documents	26,228
# of unique authors	63,695
Avg. document length	176.98
Title terms	349,410
Abstract terms	4,292,509

Author topical similarity measure

The author topical similarity measure is adopted from the topic-based author relatedness measure proposed by Lu and Wolfram (2012). The measure uses the cosine similarity between Author-Topic vectors from the training results of the Author-Topic modelling as the topical similarity between authors. Given the topic features of the Author-Topic modelling,

the author topical similarity measure is able to identify topical similarity even when the terms do not match. As is the case in any other probabilistic model, the Author-Topic modelling does not work well for authors with a limited number of publications. To ensure the quality of the topical similarity measure, we focused on authors with at least 10 publications in the Psychology subset. Higher cutoff values for the number of papers resulted in smaller numbers of authors for comparison. The cutoff of 10 papers resulted in 125 authors and 7750 author pairs. Table 2 provides descriptive statistics of the author topical similarities between the 125 prolific authors in the psychology subset.

Table 2. Descriptive statistics of the author topical similarity values (author publication count ≥ 10).

Measure	*Value*
# of author pairs	7,750
Mean	0.108
Standard deviation	0.166
Minimum	0.003
Maximum	0.997
Median	0.049

The similarity measures for the 7,750 author pairs were mapped using UCINET 6.0/NetDraw 2.1 network analysis software (https://sites.google.com/site/ucinetsoftware/home; Borgatti, Everett, & Freeman, 2002) and compared to a co-authorship map for the same authors using the data from the Elsevier Psychology dataset. The software allows the strength of ties between nodes to be represented by line thickness. Because each author topical similarity pair had essentially a non-zero similarity, the mapping of all possible author pairs resulted in an incomprehensible map filled with edges. Another advantageous feature of the software is that the display of edges may be controlled using a cutoff value. To allow the stronger relationships to be represented on the map, a similarity cutoff value of 0.5 was selected. The use of a cutoff value did not remove any data in the similarity calculation. It affected only the display of edges between author pairs by removing the edges for author similarity values below the cutoff value. Other cutoff values could have also been selected based on the strength of similarity sought. The 0.5 cutoff value resulted in 10 of the 125 authors not being included in the generated map. A co-authorship map of the Psychology authors was also generated from the Elsevier data and served as a comparison for similarity using a more commonly used measure of author similarity. There were far fewer co-authorship pairs generated from the dataset resulting in a much larger number of authors being excluded from the map because there were no collaborations present in the dataset to be represented in the map. Also, because the 0.5 cutoff value excluded 10 of the 125 authors, the same 115 authors were included in the co-authorship map. The map for the author topical similarity pairings and co-authorship relationships were compared visually for common groupings and differences.

Sampling and other data collection

To explore how the author topical similarity measure compares to other measures of similarity, a stratified random sample of 30 author pairs that spans the full range of author similarity measures was compared. Three pairs of authors were selected from each 0.1 similarity level stratum. The author topical similarity measure for each of these author pairs was compared to more commonly used similarity assessment measures including co-authorship, co-citation and mutual citations by the author pairs. The Elsevier dataset did not provide citation data and the authors did not have access to Elsevier Scopus. Citations between each author and co-citations were collected manually using Thomson Reuters Web of Science (WoS). Co-authorship data from WoS was also

incorporated because it included possible additional co-authored publications beyond those included in the Elsevier dataset. Nonparametric correlation outcomes were calculated for each measure due to the skewed distribution of the data.

Results

A histogram of the distribution of calculated author topical similarity values appears in Figure 1. Note that a logarithmic scale is used due to the large number of low similarity values. Approximately 25.7% of the similarity values exceed 0.1, and only 4.4% are above 0.5, indicating that high similarity measures may provide good discriminative capacity in distinguishing between author pairs with high and low levels of relatedness.

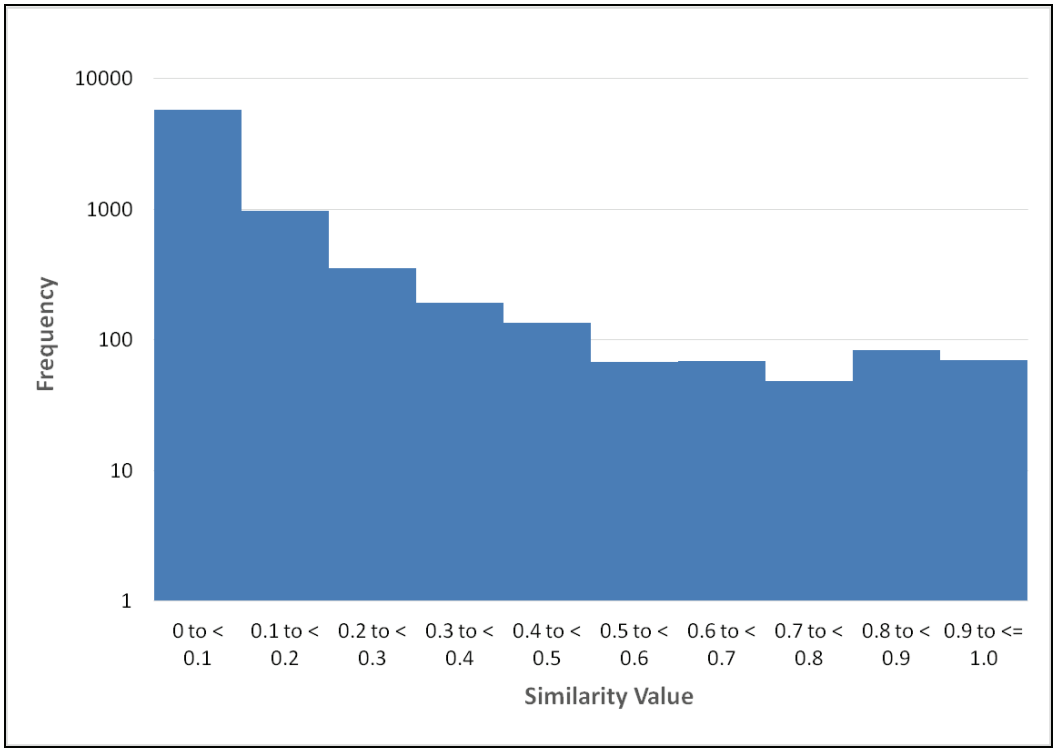


Figure 1. Histogram of calculated author topical similarity values

The UCINET map of the author topical similarity pairings with a 0.5 cutoff value appears in Figure 2. The node sizes and colours highlight comparable numbers of edges where the author similarity values are greater than 0.5. One can see that distinctive clusters of author groups are formed, with two relatively large clusters, a third mid-sized cluster and three smaller clusters with several authors. The large cluster on the right side of the map reveals an author, "Leino-Kilpi H.", who topically serves as a bridge between two parts of the cluster. The topical connection of this author to others in the cluster is missing in the co-authorship maps below due to a lack of collaboration evident in the dataset. The largest node with the greatest number of edges, "Keser H." near the centre of the large cluster to the left, indicates a high level of similarity with a large number of surrounding authors, which is also reflected in Figure 4 below.

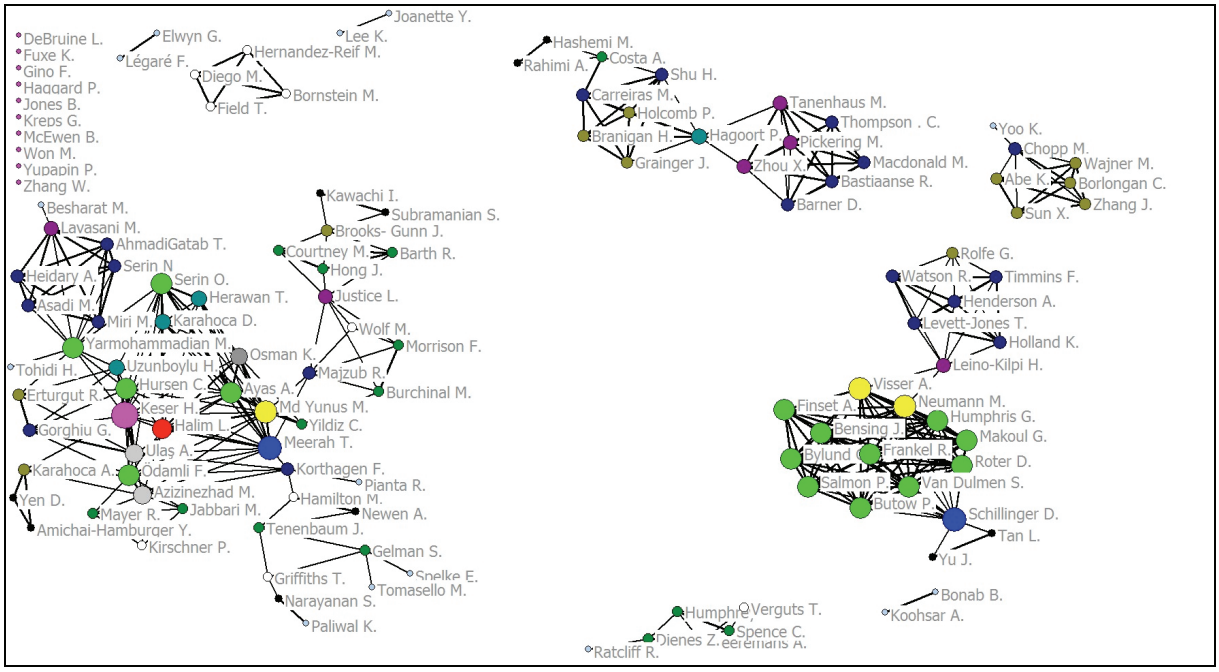


Figure 2. Author topical similarity map (similarity cutoff = 0.5).

Figure 3 summarizes the author collaboration map for the Psychology authors. One can immediately see one drawback of using co-authorship only to assess author relatedness. Fifty-six of the 125 authors were excluded because they did not collaborate with any of the other authors in the dataset. Those connections that do exist are much more limited than for the author topicality similarity outcomes, with two larger clusters and many smaller groups of two to six authors. The members of the two largest clusters in Figure 3 are almost identical to the two largest clusters in Figure 2, but represent only a fraction of the authors that appear in the Figure 2 clusters. Only three of the 10 authors excluded in Figure 2 are included in Figure 3, indicating that these three authors had no topical similarity values above 0.5 with the other authors.

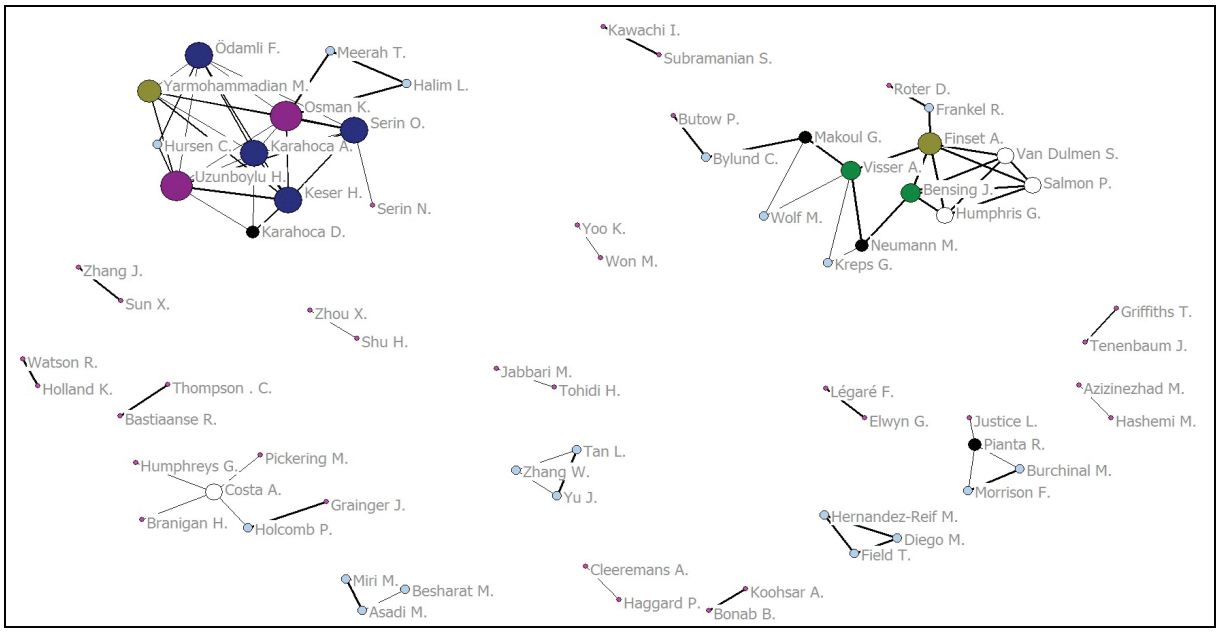


Figure 3. Co-authorship map for all author pairs

The map in Figure 2 shows relationships only for authors with a topical similarity of greater than 0.5. Figure 3 does not take into account the topical similarity of authors. Figure 4

provides the co-authorship map that includes only author pairs with a topical similarity of greater than 0.5. This eliminates a further 21 authors (77 total) from inclusion on the map. It is essentially the same map as Figure 3 flipped along the horizontal axis and, but with fewer edges arising from the removal of the additional authors.

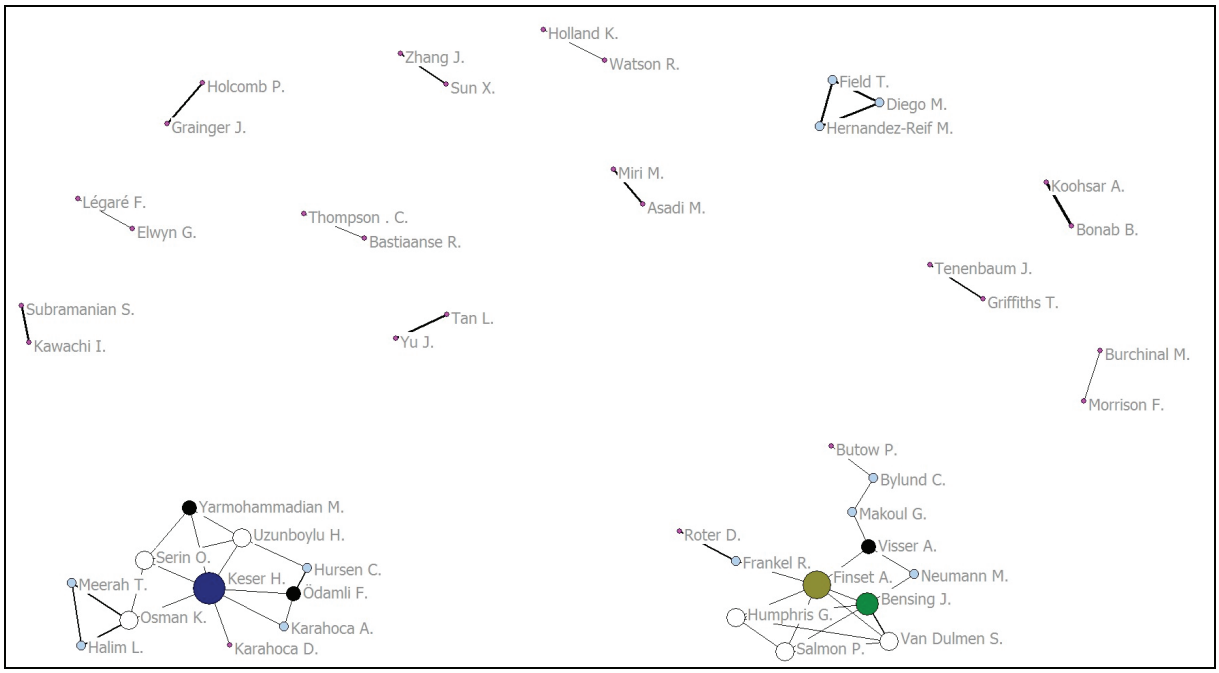


Figure 4. Co-authorship map for author pairs (Author topical similarity cutoff of 0.5)

To examine how the author topical similarity measure correlates to other author relatedness measures, 30 pairs of authors were randomly selected as described above. Outcomes for the author topical similarity, co-authorship, mutual author citing and co-citation values were compared using Spearman's rho nonparametric correlation coefficients (Table 3). There are significant, mid-level correlations observed between the author topical similarity measure and co-authorship for both the Elsevier and WoS data, as well as the mutual citing data for each author. However, there is not a significant correlation with the co-citation counts from WoS. Due to the lack of co-authorship observed for the selected author pairs for author topical similarity values below 0.5, the correlations were also run and included in Table 3 using the 15 similarity values above 0.5 (High) and the 15 values below 0.5 (Low). The positive correlations remained for high similarity author pairs but were not significant for low similarity author pairs.

Discussion

Outcomes of the author topical similarity measure provide a richer method by which author relationships may be mapped and assessed. Unlike co-authorship, direct citation and co-citation networks, where a linkage is created only through collaboration or citation behaviours. The lack of collaboration or citation does not indicate that there is no relationship between two authors; it may simply indicate that the research community has not yet recognized such a relationship. This is most evident when comparing the resulting edges based on author topical similarity and co-authorship. Even when limited to author topical similarity values of greater than 0.5, representing only 4.4% of all possible network connections, the resulting network is rich and demonstrates clusters of author relationships. The richness of the linkages in the resulting network may also be controlled by setting different cutoff values for the author topical similarity. The co-authorship map, conversely, is much sparser and only reveals explicit relationships. The relatively high correlation measure

implies that the author topical similarity measure may serve as a good predictor of existing collaboration. This is more evident for authors with higher topical similarities. Although one would expect there to be a high correlation between collaborating authors, in the Author-Topic model each word is generated from each author according to the author's profile, modelled as a distribution of topics. So, even though collaborating authors tend to be more similar, they may still be generating different words in the titles and abstracts. Excluding co-authored papers in these cases for topic modelling may be attempted, but this could result in less reliable outcomes if the majority of the text on which the models are based is removed.

Table 3. Spearman's rho correlation outcomes for author relatedness measures

		Author Topical Similarity	*Co-authorship Elsevier*	*Co-authorship WoS*	*A Cites B WoS*	*B Cites A WoS*	*Co-citation WoS*
Author Topical Similarity	All	1	.568**	.660**	.452*	.445*	.255
	High	1	.710**	.762**	.694**	.691**	.311
	Low	1	NA	NA	.141	.099	.373
Co-authorship Elsevier	All		1	.816**	.414*	.490**	.415*
	High		1	.812**	.531*	.607*	.573*
	Low		NA	NA	NA	NA	NA
Co-authorship WoS	All			1	.472**	.374*	.336
	High			1	.583*	.398	.492
	Low			NA	NA	NA	NA
A Cites B WoS	All				1	.669**	.587**
	High				1	.812**	.505
	Low				1	.492	.728**
B Cites A WoS	All					1	.536**
	High					1	.451
	Low					1	.650**
Co-citation WoS	All						1
	High						1
	Low						1

** Correlation is significant at the 0.01 level (2-tailed).

* Correlation is significant at the 0.05 level (2-tailed).

In the absence of existing collaborations, high author similarity values could serve as an indicator for possible future collaborations. We recognize that the motivations for collaboration are complex and go beyond authors having similar interests. Collaboration may also be prompted by the complementary areas of expertise collaborators bring, which would not be reflected using topic modelling techniques alone. Still, the similarity measure may be used to identify research "birds of a feather" that may not be evident using similarity measures based on collaboration or citation data.

In answer to the research questions posed at the beginning of this paper: 1) mapping of author relatedness based on author topical similarity can reveal a richer network of relationships between authors not evident through a more traditional relationship assessment based on co-authorship and can identify topical bridges; 2) co-authorship, co-citation and mutual citation between authors are significantly correlated, in particular for authors with high topical similarity, so authors with similar topical interests may be more likely to collaborate or cite each other; 3) high author topical similarity values can serve as a reasonably accurate predictor of co-authorship and mutual citation, but not of co-citation activity. The lack of a significant correlation between author topical similarity and co-citation provides evidence that

the topical similarity measure offers a different perspective on author relationships that complements the more traditional co-citation approach. The significant correlations observed between the author topicality similarity and other citation and co-authorship measures indicate that topicality may be a weak to moderately strong predictor of these other more traditional measures for authors with high topical similarity. This positive correlation between author topicality and co-authorship is not unexpected given that co-authored publications would result in more similar topical assignments.

The findings of this study have implications for author relatedness assessment. As the author topical similarity measure does not depend on collaboration or citation behaviour, it can serve as an alternative author relatedness measure where there is a lack of collaboration or citation connections. Even if the collaboration and citation connections exist, the topical similarity measure can provide complementary evidence of relatedness from the content perspective. In addition, the significant correlations between author topical similarity and collaboration shed light on recent developments in predicting and recommending collaborations. Most existing methods for predicting and recommending collaborations are based on the topological features of collaboration networks (Yan & Guns, 2014). The level of correlations between topical similarity and collaboration, particularly for authors with high similarity, provide strong evidence of including content-based predictors for this problem.

Topic modelling offers the ability to reveal relationships between authors that may not be evident through more traditional methods of similarity assessment, but it does have its limitations. The computational overhead associated with topic-based author relatedness modelling is more substantial than for citation and collaboration-based data. Also there must be a sufficient body of text to train the topic model and to accurately represent author relationships; therefore, this method may not be suitable for authors with a more modest publication record. In this case, analysis using citation-based methods may be more fruitful. Other limitations arise from the dataset itself. In identifying works attributable to an author, we have relied on the supplied Scopus author identifier. We recognize that author name disambiguation, regardless of the method used, may not be 100% accurate. In addition, the present study has limited itself to data from a single discipline. Furthermore, the dataset itself was not complete for the discipline of Psychology, but rather a subset. We cannot conclude that the outcomes for other disciplines will be similar. Outcomes would depend also on the collaboration traditions and citing behaviours of those disciplines. The computational overhead and limited ability for topic modelling to be able to produce meaningful topics with multidisciplinary datasets may limit the application of this approach beyond the disciplinary level.

Conclusions

Author topical similarity provides a novel way to assess author relatedness that complements existing methods based on co-authorship, direct citation or co-citation. While other methods require an existing form of connection based on collaboration or citations, author topical similarity assesses author relatedness based on the language used by the authors themselves. The small percentage of author pairs with high similarity values indicates that the measure is discriminating in the assessment of author relatedness. The present study has demonstrated how author relatedness based on topic modelling can provide a richer method to assess how closely related authors' research contributions are. Although significantly correlated with co-authorship and direct citation measures, author topical similarity between authors was not found to be significantly correlated with co-citations, which has been commonly used to assess author relatedness. Author topical similarity outcomes may serve as a reasonably accurate predictor of existing collaborations between authors, or an indicator of potential future collaborators in the absence of existing collaboration. Future research may investigate

how author topical similarity measures compare to other existing author relatedness measures for other disciplinary areas including the humanities and sciences, where collaboration and citation patterns may differ.

Acknowledgments

The authors would like to thank Elsevier, Inc. for providing access to the dataset used to conduct this study.

References

Blei, D., & Lafferty, J. (2006). Correlated topic models. *Advances in Neural Information Processing Systems*, *18*, 147.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet Allocation. *Journal of Machine Learning Research*, *3*, 993–1022.

Borgatti, S. P., Everett, M. G., & Freeman, L. C. (2002). UCINET for Windows: Software for social network analysis. Retrieved from https://www.google.com/url?sa=t&rct=j&q=&esrc=s&source=web&cd=6&ved=0CEMQFjAF&url=https%3A%2F%2Fwww.soc.umn.edu%2F~knoke%2Fpages%2FUCINET_6_User%2527s_Guide.doc&ei=WqatVLfOKtP3ggTgq4OIAg&usg=AFQjCNF_1umvC9bg07zqAb969AG7WJX5qw&cad=rja

Börner, K., Chen, C., & Boyack, K. W. (2003). Visualizing knowledge domains. *Annual Review of Information Science and Technology*, *37*(1), 179–255.

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.

Glänzel, W. & Schubert, A. (2005). Analysing scientific networks through co-authorship. In *Handbook of Quantitative Science and Technology Research* (pp. 257-276). Netherlands: Springer.

Hofmann, T. (1999). Probabilistic Latent Semantic Indexing. In Proceedings of the 22Nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (pp. 50–57). New York, NY, USA: ACM.

Hu, Y., Boyd-Graber, J., Satinoff, B., & Smith, A. (2014). Interactive topic modeling. *Machine Learning*, *95*(3), 423–469.

Jacobs, D., & Wolfram, D. (2014). Exploring author similarity using citing discipline analysis. In *Proceedings of the Annual Conference of CAIS/ Actes du congrès annuel de l'ACSI*. Retrieved from: http://www.cais-acsi.ca/ojs/index.php/cais/article/download/892/812.

Jeong, Y. K., Song, M., & Ding, Y. (2014). Content-based author co-citation analysis. *Journal of Informetrics*, *8*(1), 197-211.

Law, J., & Whittaker, J. (1992). Mapping acidification research: A test of the co-word method. *Scientometrics*, *23*(3), 417–461.

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.

Mcauliffe, J. D., & Blei, D. M. (2008). Supervised topic models. In *Advances in Neural Information Processing Systems* (pp. 121–128). Retrieved from http://papers.nips.cc/paper/3328-supervised-topic-models

Rosen-Zvi, M., Chemudugunta, C., Griffiths, T., Smyth, P., & Steyvers, M. (2010). Learning author-topic models from text corpora. *ACM Transactions on Information Systems*, *28*(1), 1-38.Tang, J., Zhang, J., Yao, L., Li, J., Zhang, L., & Su, Z. (2008). ArnetMiner: Extraction and Mining of Academic Social Networks. In *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 990–998). New York, NY, USA: ACM.

Tang, J., Jin, R., & Zhang, J. (2008). A topic modeling approach and its integration into the random walk framework for academic search. In F. Giannotti, D. Gunopulos, F. Turini, C. Zaniolo, N. Ramakrishnan, & X. Wu (Eds.), *Eighth IEEE International Conference on Data Mining ICDM'08*. (pp. 1055-1060). IEEE.

White, H.D., & McCain, K.W. (1998). Visualizing a discipline: An author co-citation analysis of information science, 1972-1995. *Journal of the American Society for Information Science, 49*(4)*,* 327-355.

Yan, E., & Guns, R. (2014). Predicting and recommending collaborations: An author-, institution-, and country-level analysis. *Journal of Informetrics, 8*(2)*,* 295-309.

Zhao, D., & Strotmann, A. (2008). Evolution of research activities and intellectual influences in Information Science 1996-2005: Introducing author bibliographic-coupling analysis. *Journal of the American Society for Information Science and Technology*, *59*(13), 2070–2086.

Zhao, D., & Strotmann, A. (2014). The knowledge base and research front of information science 2006–2010: An author cocitation and bibliographic coupling analysis. *Journal of the Association for Information Science and Technology*, *65*(5), 995-1006.

Publication Rates in 192 Research Fields of the Hard Sciences

Ciriaco Andrea D'Angelo[1] and Giovanni Abramo[2]

[1] *dangelo@dii.uniroma2.it*
Department of Engineering and Management
University of Rome "Tor Vergata" – Italy, Via del Politecnico 1, 00133 Rome (Italy)

[2] *giovanni.abramo@uniroma2.it*
Laboratory for Studies of Research and Technology Transfer, Institute for System Analysis and Computer Science (IASI-CNR), National Research Council of Italy, Via dei Taurini 19, 00185 Rome (Italy)

Abstract
Bibliometricians are aware that the citation behavior of scientists varies across fields, and for this they carefully normalize citations by field. They are also aware of the different publication intensities across fields. This imposes that the research performance of a scientist must be compared with that of their colleagues in the same field. Every comparison of scientists in different fields should be preceded by the normalization of the performances, and the same holds for comparing multidisciplinary organizational units. If the Web of Science recognizes 251 subject categories, there should be a somewhat similar number of research fields for the classification of the scientists. The Italian academic system is quite unique in providing a classification of professors, into 370 fields, 192 of them in the hard sciences. In this work we measure the descriptive statistics on annual publication (full and fractional counting) by Italian academics in each of the 192 hard science fields. These statistics help recognize the extent of distortion from failing to normalize the research performance of scientists based in different fields. They could also serve as scaling factors for avoiding distortion in rankings, including in other nations.

Conference Topic
Methods and techniques

Introduction
The purpose of bibliometrics is to provide continuously better support for the policy-makers and administrators of research institutions, in the achievement of their specific objectives, through the provision of methods and indicators for the evaluation of performance that are themselves always more accurate, robust, reliable and functional. The principle obstacle to bibliometrics is the insufficiency of the data to meet such high standards. The practitioner is thus forced to resort to proxies in measurement, which cause varying degrees of distortion in the results.

Research organizations are likened to other productive organizations, but where the product is new knowledge, rather than some other good or service. An organization's performance is then better than that of another one if, at parity of resources, it produces more knowledge or if, at parity of output, it consumes less resources. It is the shortage of information on inputs (production factors) that presents the greatest problem to bibliometricians. The production factors are labor and capital. Capital embeds all those resources other than labor (facilities, technical instruments, materials, databases, etc.). When we wish to measure labor productivity we must thus normalize for capital. But who can really know the financial and technical resources available to all the different institutions, departments, and then individual researchers? The bibliometrician also frequently lacks information on the realities of labor, due to the absence of databases on the researchers, and on their institutional, discipline and field affiliations.

Given these obstacles, practitioners often use indicators that do not relate output to input. This means they produce ranking lists that are highly size-dependent. At that point we cannot

know what part of an organization's or nation's rank arises from its performance or is due to size. Examples of this are the CWTS Leiden[1] and SCImago[2] lists, which rank universities by publications and fractional publications. Others have proposed indicators that attempt to get around the problems by relating the impact or excellence of research not to input, but rather to the output itself. Examples of this are the "new crown indicator" (Waltman et al., 2011), which measures the average impact per publication, or the "proportion of highly-cited articles to total publications" (Waltman et al., 2012). However, with this type of indicator, even when the output of the scientist increases, other factors remaining equal, his or her performance could still decrease: a paradox and a violation of the fundamental principle of the measure of efficiency.

In those cases where an indicator does relate output to input, it is still often applied at levels of organizational aggregation that are too high, ignoring the differing intensity of publication across fields. Bibliometricians have been aware of this problem for many years (Butler, 2007; Moed et al., 1985; Garfield, 1979), and are also aware of the distortion that afflicts the resulting aggregate rankings (Abramo, D'Angelo, & Di Costa, 2008). However the task of finer aggregation is difficult to solve without a database that classifies the researchers by field of research. Where they exist, such databases are maintained at central levels. Apart from the Italian one[3], maintained by the Italian Ministry of Education, Universities and Research (MIUR), the only other large-scale one we are aware of is the Norwegian Research Personnel Register[4] compiled by the Nordic Institute for Studies in Innovation, Research and Education (NIFU).

The NIFU system classifies scientists in 58 scientific fields grouped in five main domains. Perhaps the lower number of scientists in Norway works against finer classification: in fact comparing the performance of small numbers of researchers per field creates serious problems of significance. However, on the other hand, the Web of Science (WoS) identifies a full 251 subject categories for the classification of journals. And if there are this many fields for classifying scientific journals, there must be at least that many fields for classifying scientific work, and the scientists. In smaller nations or emerging economies we could expect to see fewer number of these fields present, since research structures will be unable to deal with all the areas, and we would expect to see research in more concentrated fields. However, in larger, developed countries we can expect to see the full spectrum of research fields. In fact in Italy the MIUR manages a system for the classification of all professors into a total of 370 "scientific disciplinary sectors" (SDSs).[5] Each professor belongs to one and only one of the SDSs, which are grouped into 14 university disciplinary areas (UDAs). Further, 192 of the SDSs from 9 of the UDAs fall in the so-called hard sciences. In the following we refer to these SDS by their code or acronym.[6] These 192 SDSs compare to the 176 WoS subject categories identified in the JCR-Science Citation Index (see the Annex 1[7] for a conversion of SDSs to WoS subject categories.

As noted above, the lack of field classification of scientists means that measures of research performance will inevitably be affected by distortions in rankings, due to the different intensity of publication across fields. The higher the level of aggregation, the stronger these distortions become. The corollary is that, rising to international levels, it has been impossible

[1] http://www.leidenranking.com/ranking/2014, last accessed on April 8, 2015.
[2] http://www.scimagoir.com/research.php, last accessed on April 8, 2015.
[3] http://cercauniversita.cineca.it/php5/docenti/cerca.php, last accessed on April 8, 2015.
[4] http://www.nifu.no/en/statistikk/databaser-og-registre/4897-2/ last accessed on April 8, 2015.
[5] The complete list is accessible on attiministeriali.miur.it/UserFiles/115.htm, last accessed April 8, 2015.
[6] The full names can be found in www.iasi.cnr.it/laboratoriortt/TESTI/Altro/ISSI-ANNEX%202_P.pdf, last accessed on April 8, 2015
[7] www.iasi.cnr.it/laboratoriortt/TESTI/Altro/ISSI-ANNEX1.pdf, last accessed on April 8, 2015

to correctly compare institutional or national research performance.

To date, in fact there is no international standard for the classification of scientists. Thus in this work we provide our colleagues and practitioners with descriptive statistics on yearly publications (both full and fractional counting) of Italian academics in each of the 192 hard science SDSs. Our intention is that these statistics might first permit recognition of the extent of distortions that occur when evaluations compare the research performance of scientists within the same discipline, but in different fields. For those nations lacking databases of researchers by field, our statistics could also serve as normalization factors, serving to reduce the distortions when comparing research performance of individuals, groups or entire research organizations.

Data and Methods

In the study we measure "publication rates" in 192 SDSs, meaning average yearly publications of individual scientists, over the period 2009-2013.[8] Data on Italian academics are extracted from the official database maintained by the MIUR. The database indexes the name, academic rank, affiliation, and SDS of all academics in Italian universities. At 31/12/2013 the entire Italian university population consisted of 56,600 professors employed in 96 universities, which are authorized by the MIUR to grant legally recognized degrees. 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. For reasons of robustness, we thus examine only the nine UDAs that deal with the hard sciences,[9] including a total of 192 SDSs. Furthermore, again for reasons of robustness, we calculate the yearly average publication rates only of those professors who have been on staff for at least three years over the observed period.

Table 1. Dataset for the analysis: number of fields (SDSs), universities, research staff and WoS publications in each UDA under investigation

UDA	SDS	Universities	Research staff	Publications*
Mathematics and computer science	10	72	2,930	16,262
Physics	8	65	2,003	22,597
Chemistry	12	60	2,701	26,054
Earth sciences	12	49	974	6,066
Biology	19	67	4,423	34,406
Medicine	50	65	8,998	72,661
Agricultural and veterinary sciences	30	56	2,820	14,951
Civil engineering	9	54	1,394	7,462
Industrial and information engineering	42	73	4,791	40,572
Total	192	86	31,034	207,132†

* Figures refer to publications authored by at least one professor pertaining to the UDA.

† Total is less than the sum of the column data due to double counts of publications co-authored by researchers pertaining to SDSs of more than one UDA.

Publication data are drawn from the Italian Observatory of Public Research (ORP), a database developed and maintained by the authors and derived under license from the WoS. Beginning from the raw data of Italian publications[10] indexed in WoS-ORP, we apply a complex

[8] For the most appropriate publication period to be observed see Abramo et al. (2012b).

[9] Mathematics and computer sciences; Physics; Chemistry; Earth sciences; Biology; Medicine; Agricultural and veterinary sciences; Civil engineering; Industrial and information engineering.

[10] We exclude those document types that cannot be strictly considered as true research products, such as editorial material, meeting abstracts, replies to letters, etc.

algorithm for disambiguation of the true identity of the authors and their institutional affiliations (for details see D'Angelo et al., 2011). Each publication is attributed to the university professors that authored it, with a harmonic average of precision and recall (F-measure) equal to 96 (error of 4%). We further reduce this error by manual disambiguation. Because each professor belongs to one and only one SDS, we can then calculate the distribution of annual publication rates and the relevant descriptive statistics in each SDS.

The dataset for the analysis includes 31,034 professors, employed in 86 universities, authoring over 200,000 WoS publications, sorted in the UDAs as shown in Table 1.

Research projects frequently involve a team of researchers, a fact revealed in the co-authorship of publications. Various performance measures account for the fractional contributions of single co-authors 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 the ordering of the byline. The conventions on the order of authors for scientific papers differ across fields (Pontille, 2004; RIN, 2009), thus in the current study, the fractional contribution of the individuals is weighted accordingly.

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, particularly in the life sciences. For these disciplines, 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.[11] Failure to account for the number and position of authors in the byline would result in notable ranking differences, both at the individual level (Abramo, D'Angelo & Rosati, 2013a) and at the institution level (Abramo, D'Angelo & Rosati, 2013b).

Applying the above conventions, for each of the 192 SDS we will provide descriptive statistics on the intensity of annual publication: referred to as P for full counting and FP for fractional counting. We then examine further statistics on P and FP for the SDSs included in each UDA.

Results

Publication rates of professors in a specific field

The publication intensity of professors in a given field is known to be particularly skewed, with a small percentage of individuals authoring a large share of the total papers, and the others authoring a small share (Egghe, 2005; Kyvik, 1989; Lotka, 1926). Figure 1 provides the example of the field of Organic chemistry (SDS CHIM/06), showing the distribution of the average number of publications per year over the period under examination, for each of the 554 professors in the SDS. The distribution fits quite well a logarithmic curve, as indicated by the particularly high value of R^2 (0.974). Here, 10% of the professors have produced on average less than one publication per year, and six were totally unproductive. On the opposite front, we find 20 professors with over 10 publications per year, and one absolute outlier with 25.

The box plot (right side of Figure 1) refers to the same distribution. It shows a median of 3 publications per year and an interquartile range (difference between third and first quartile) of

[11] 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.

2.6. It also brings out the presence of 30 outliers: hyper-productive professors with a performance that exceeds that of the third quartile by over 1.5 times the interquartile difference.

The distribution of frequencies by class of publication rates (Figure 2) shows a mode between 2 and 3 publications annually and a particularly long right tail, with a final peak for the hyper-productive professors.

The distribution of the average yearly publications measured by fractional counting (FP) shows a very similar situation: in Figure 3 the right tail is actually longer than that for only full counting (Figure 2).

The distributions seen for SDS CHIM/06 show structural elements that recur in the analyses of the other 191 SDSs. Most obvious is the skewness, although there are some interesting exceptions, for example as in VET/04 (Inspection of food products of animal origin). The 77 professors of this SDS have a publication rate that is almost uniform, as illustrated in Figure 4.

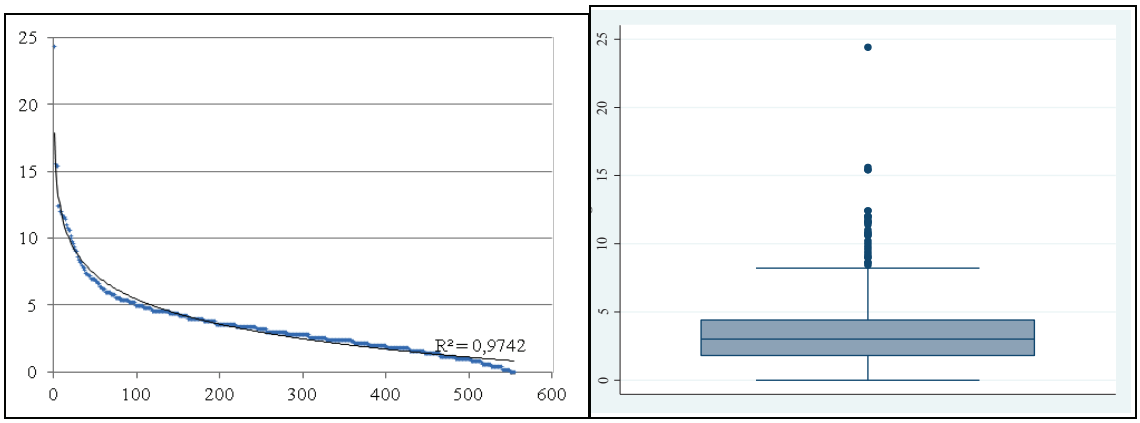


Figure 1. Distribution and box plot of annual publication rate P (full counting, 2009-2013) for 554 Italian professors in Organic chemistry (CHIM/06).

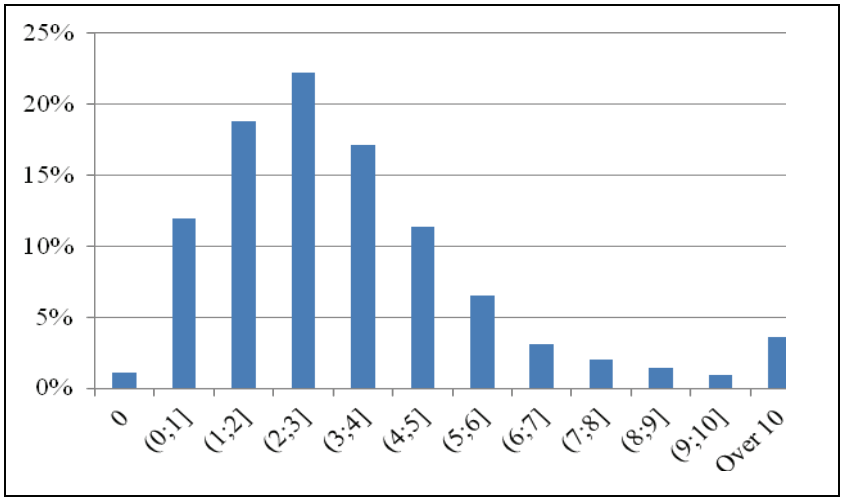


Figure 2. Frequency distribution for classes of annual publication rate P (2009-2013) for the 554 Italian professors in CHIM/06.

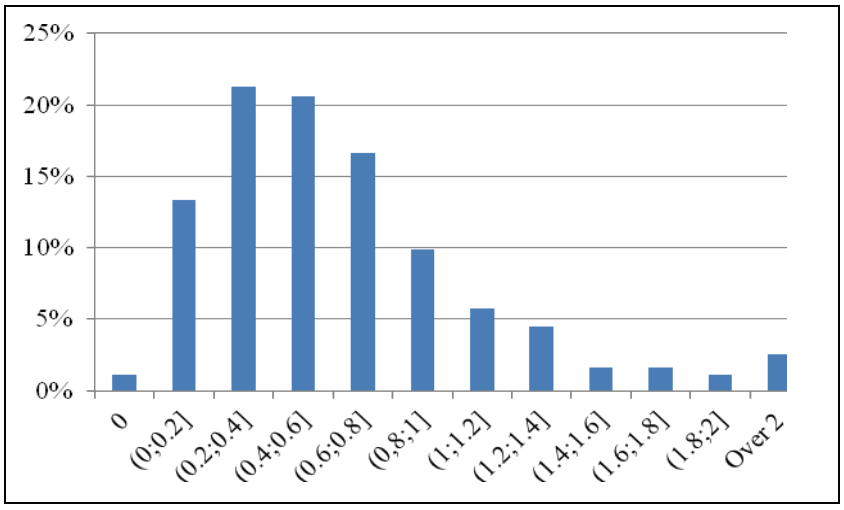


Figure 3. Frequency distribution for classes of annual publication rate FP (fractional counting, 2009-2013) for the 554 Italian professors in CHIM/06.

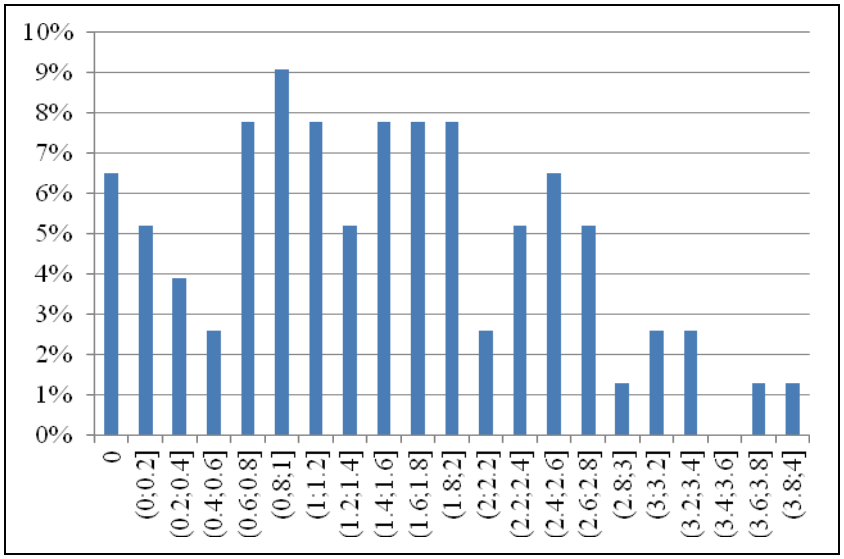


Figure 4. Frequency distribution for classes of annual publication rate P (2009-2013) for Italian professors in Inspection of food products of animal origin (VET/04).

Publication rates of fields within a discipline

As with the two examples above (CHIM/06 and VET/04), the publication rates in the various SDSs are never superimposable. Thus the calculation of the descriptive statistics for the SDSs provides useful benchmarks for the professors that work in them. Table 2 provides the statistics for all the SDSs in the Earth sciences discipline.

This UDA consists of a total of 12 SDSs with very different sizes in terms of national research staff, from a minimum of 17 professors in Applied geophysics (GEO/12) to a maximum of 137, in Palaeontology and palaeoecology (GEO/02). The intensity of publication is structurally very different. In Stratigraphic and sedimentological geology (GEO/03) only 2.2% of the professors (2 of 92) did not produce any publications over the five-year period under examination. On the opposite front there are 19 unproductive professors among the 121 of Physical geography and geomorphology (GEO/05), or 15.7% of the total. This SDS also registers the lowest average annual rate of publication, at 1.12 per year, followed by Structural geology (GEO/04), GEO/02 and Geophysics of solid earth GEO/11 (1.44, 1.48 and 1.49, respectively). In half the SDSs there is an average intensity of publication of 2 per year,

with a peak in Applied geology GEO/06 (3.09). Clearly, among all those of the UDA, this SDS has the greatest publication rate: the distribution of the performances shows all values in the highest quartiles. The top 25% of professors (3rd quartile) produce on average more than 4 publications per year, with the absolute record being a professor who produces almost 18. The dispersion of the performances in all the SDSs, indicated by the variation coefficients in the last column of Table 2, results as greatest in GEO/03 and GEO/05, where the coefficient is above 1.

The analyses of the distributions for fractional counting of the publication rate (FP) (Table 3) provide a picture similar to that for full counting. The average intensity of collaboration evidently does not vary in a substantial way between the SDSs, and thus the differential of publication rates between the SDSs does not vary in going from a full counting approach to fractional counting.

Table 2. Descriptive statistics for intensity of annual publication rate P (2009-2013) for the SDSs of Earth sciences.

SDS	Research staff	Unproductive	I quartile	Median	III quartile	Max	Average	Std dev.	Variat. coeff.
GEO/01	93	3.2%	0.8	1.6	2.2	8	1.76	1.40	0.80
GEO/02	137	7.3%	0.6	1	2.20	6.4	1.48	1.25	0.84
GEO/03	92	2.2%	1	1.8	2.8	22	2.40	2.69	1.12
GEO/04	116	6.9%	0.6	1	2	4.8	1.44	1.21	0.84
GEO/05	121	15.7%	0.2	0.8	1.4	8.2	1.12	1.22	1.09
GEO/06	76	1.3%	1.55	2.6	4.05	17.8	3.09	2.51	0.81
GEO/07	82	2.4%	1	1.8	2.75	8.2	1.99	1.46	0.73
GEO/08	67	3.0%	1.3	2.4	3.5	10.6	2.69	2.03	0.75
GEO/09	63	6.3%	0.8	1.8	2.9	11.4	2.21	2.04	0.92
GEO/10	69	4.3%	1.2	1.8	2.4	10.2	2.14	1.82	0.85
GEO/11	41	2.4%	0.6	1.2	2	5.6	1.49	1.12	0.75
GEO/12	17	5.9%	0.8	1.6	2	4.6	1.75	1.34	0.77

Table 3. Descriptive statistics for intensity of annual publication rate FP (2009-2013) for the SDSs of Earth sciences

SDS	I quartile	Median	III quartile	Max	Average	Std dev.	Variat. coeff.
GEO/01	0.20	0.33	0.53	2.61	0.45	0.45	1.00
GEO/02	0.14	0.28	0.45	1.47	0.34	0.29	0.85
GEO/03	0.26	0.43	0.65	2.64	0.53	0.42	0.79
GEO/04	0.14	0.25	0.47	1.81	0.33	0.30	0.91
GEO/05	0.07	0.24	0.42	1.81	0.29	0.31	1.07
GEO/06	0.32	0.56	0.87	3.52	0.71	0.61	0.86
GEO/07	0.20	0.37	0.59	1.62	0.44	0.31	0.70
GEO/08	0.29	0.53	0.72	1.61	0.56	0.39	0.70
GEO/09	0.13	0.39	0.65	3.06	0.48	0.48	1.00
GEO/10	0.29	0.45	0.74	2.44	0.56	0.46	0.82
GEO/11	0.19	0.31	0.61	1.50	0.45	0.38	0.84
GEO/12	0.19	0.32	0.60	0.90	0.38	0.28	0.74

For the descriptive statistics of the full 192 SDSs investigated, we refer the reader to Annex 2[12] for the full counting, and to Annex 3[13] for fractional counting. Below, in Table 4, we show for each UDA the SDSs with minimum and maximum values of some of the above statistics

[12] www.iasi.cnr.it/laboratoriortt/TESTI/Altro/ISSI-ANNEX%202_P.pdf, last accessed on April 8, 2015
[13] www.iasi.cnr.it/laboratoriortt/TESTI/Altro/ISSI-ANNEX%203_FP.pdf, last accessed on April 8, 2015

of P (full counting). The data indicate substantial variability in the intensity of publication between the SDSs in all the UDAs. In Mathematics the percentage of unproductive professors varies from a minimum of 3.9% in MAT/09 (Operations research) and a maximum of 43.2% in MAT/04 (Complementary mathematics). Such substantial variations also occur in Medicine, with 1.1% unproductive professors in MED/08 (Pathological anatomy) and 45.5% in MED/02 (History of medicine). In Agricultural and veterinary sciences, VET/02 (Veterinary physiology) does not have any unproductive professors, while AGR/01 (Rural economics and valuation) registers a share of 45.5%. More contained heterogeneity in unproductive professors is seen in some other UDAs: certainly in Earth sciences, which we have already examined, but also in Biology. In this UDA the maximum incidence of unproductive professors (11.8% of the total professors) is seen in BIO/08 (Anthropology) and the minimum (1.2%) in BIO/15 (Pharmaceutical biology). The median intensity of annual publication also presents high variability between the SDSs of a UDA. In Mathematics the median ranges from 0.2 publications per year in MAT/04 (Complementary mathematics) to 1.8 in MAT/09 (Operations research). In effect the interval of variation of the median values is very substantial in almost all the UDAs. Within Industrial and information engineering, the median intensity of publication registered in ING-INF/06 (Electronic and information bioengineering) and in ING-INF/02 (Electromagnetic fields) is more than 40 times that registered in ING-IND/01 (Naval architecture). In Medicine the two extreme situations concern MED/02 (History of medicine) and MED/16 (Rheumatology): the median intensity of publication registered in the first SDS (0.2) is $1/25^{th}$ of that for the second (5.0). The differences are more contained in Chemistry (2.0 vs. 3.4), Earth sciences (0.8 vs. 2.6) and Biology (1.1 vs. 3.3). The consistency of the outliers is also significantly different between the SDSs of a given discipline. In the Mathematics UDA, the most productive professor in absolute terms is one in INF/01 (Computer science), with an average of 28.6 publications per year, against the 3.6 of the most productive professor in MAT/04 (Complementary mathematics). In Medicine, a professor in MED/24 (Urology) registers a median of 76 publications per year over the five years examined; the most prolific in MED/47 (Nursing and midwifery) has barely 1.4 publications. In Industrial and information engineering the most prolific professor of ING-IND/01 (Naval architecture) authors an average of 1.4 publications annually, against the 33.2 of the most productive in ING-IND/34 (Industrial bioengineering). Finally, Physics FIS/01 (Experimental physics) includes a professor with an average of over 100 publications per year. In effect, this SDS consists of a range of subfields, including "high energy physics", where scientists regularly author hundreds of publications together with hundreds of co-authors. In this case (but not only in this case) a more opportune benchmark could be the distribution of the publication rate under the fractional counting method. Table 5 shows, for every UDA, the SDS with minimum and maximum values of the main statistics[14] of the fractional counting distributions. We see a level of superimposability with the data of Table 4, both in terms of the SDSs featured and for the intervals of variation in the main statistics of the SDSs, for each UDA.

[14] To avoid pointless duplication, the table does not show the incidence of unproductive professors, and instead provides statistics on average publication rate.

Table 4. SDSs with Min and Max values of descriptive statistics of intensity of annual publication P (2009-2013), for all UDAs.

	Unproductive (%)		Median		Max	
UDA*	Min	Max	Min	Max	Min	Max
1	3.9 (MAT/09)	43.2 (MAT/04)	0.2 (MAT/04)	1.8 (MAT/09)	3.6 (MAT/04)	28.6 (INF/01)
2	2.1 (FIS/04)	37.5 (FIS/08)	0.2 (FIS/08)	5.6 (FIS/01)	4.4 (FIS/08)	102.2 (FIS/01)
3	0.0 (CHIM/04)	8.6 (CHIM/11)	2.0 (CHIM/11)	3.4 (CHIM/02)	7.6 (CHIM/12)	66.2 (CHIM/08)
4	1.3 (GEO/06)	15.7 (GEO/05)	0.8 (GEO/05)	2.6 (GEO/06)	4.6 (GEO/12)	22 (GEO/03)
5	1.2 (BIO/15)	11.8 (BIO/08)	1.1 (BIO/02)	3.3 (BIO/15)	6.4 (BIO/08)	37.6 (BIO/12)
6	1.1 (MED/08)	45.5 (MED/02)	0.2 (MED/02)	5.0 (MED/16)	1.4 (MED/47)	76 (MED/24)
7	0.0 (VET/02)	42.0 (AGR/01)	0.2 (AGR/01)	2.8 (VET/06)	3.2 (AGR/06)	32.6 (VET/06)
8	5.8 (ICAR/03)	29.9 (ICAR/06)	0.2 (ICAR/06)	1.6 (ICAR/03)	2.8 (ICAR/05)	21.2 (ICAR/08)
9	0.0 (ING-IND/18)	50.0 (ING-IND/01)	0.1 (ING-IND/01)	4.4 (ING-INF/02 and ING-INF/06)	1.4 (ING-IND/01)	33.2 (ING-IND/34)

* 1 = Mathematics and computer sciences; 2 = Physics; 3 = Chemistry; 4 = Earth sciences; 5 = Biology; 6 = Medicine; 7 = Agricultural and veterinary sciences; 8 = Civil engineering; 9 = Industrial and information engineering

Table 5. SDSs with Min and Max values of descriptive statistics of intensity of annual publication FP (2009-2013), for all UDAs.

	Median		Average		Max	
UDA*	Min	Max	Min	Max	Min	Max
1	0.10 (MAT/04)	0.55 (MAT/09)	0.16 (MAT/04)	0.70 (MAT/07)	1.00 (MAT/04)	6.47 (MAT/02)
2	0.07 (FIS/08)	0.74 (FIS/03)	0.20 (FIS/08)	0.96 (FIS/03)	0.80 (FIS/08)	13.74 (FIS/03)
3	0.35 (CHIM/12)	0.70 (CHIM/02)	0.58 (CHIM/12)	0.83 (CHIM/02)	2.38 (CHIM/12)	17.60 (CHIM/08)
4	0.24 (GEO/05)	0.56 (GEO/06)	0.29 (GEO/05)	0.71 (GEO/06)	0.90 (GEO/12)	3.52 (GEO/06)
5	0.24 (BIO/08)	0.58 (BIO/15)	0.32 (BIO/08)	0.85 (BIO/15)	1.04 (BIO/08)	10.50 (BIO/12)
6	0.01 (MED/02)	0.84 (MED/16)	0.08 (MED/47)	1.18 (MED/16)	0.19 (MED/47)	13.28 (MED/11)
7	0.04 (AGR/01)	0.60 (AGR/15)	0.14 (AGR/01)	0.78 (VET/06)	0.65 (AGR/06)	9.14 (VET/06)
8	0.10 (ICAR/06)	0.48 (ICAR/08)	0.17 (ICAR/06)	0.73 (ICAR/08)	1.27 (ICAR/05)	6.85 (ICAR/08)
9	0.03 (ING-IND/01)	1.08 (ING-INF/02)	0.10 (ING-IND/01)	1.28 (ING-INF/02)	0.54 (ING-IND/02)	9.18 (ING-IND/19)

* 1 = Mathematics and computer sciences; 2 = Physics; 3 = Chemistry; 4 = Earth sciences; 5 = Biology; 6 = Medicine; 7 = Agricultural and veterinary sciences; 8 = Civil engineering; 9 = Industrial and information engineering

Conclusions

The great majority of the bibliometric indicators and the relative rankings lack fine-grained normalization of performance to the field to which the scientists belong. While bibliometricians intelligently field-normalize citations to account for the different citation behaviors across fields, they often close an eye when it comes to accounting for the different intensity of publication. At most they distinguish scientists as belonging to a few large disciplines, which cannot be sufficient if we accept the WoS as a true characterization, where scientific work is distinguished in 251 subject categories. Why would we normalize the citations for these 251 subject categories but then the scientists' performance for only a few disciplines? The answer is simple: in most cases the bibliometricians lack information about the field of research of each scientist under observation. Even at the national level the challenge of identifying the scientist's field is daunting, let alone for the task of international comparison.

Taking advantage of a particular feature of the Italian academic system, in this work we have provided descriptive statistics on the yearly publication rates of all Italian professors (over 30,000) in each of the 192 hard sciences fields, with both full and fractional counting method. Although the dataset refers to a specific nation, the very substantial size and the fine-grained field stratification certainly make it a useful reference system for the comparative evaluation of scientists in all the world. The only condition is that scholars recognize in which field of the Italian system the core of their scientific production falls. To this aim, in the Appendix, we have provided the reader with a conversion table, which establishes a link between SDSs and WoS subject categories, based on incidence of publications authored by Italian academics. Through this link, scientists outside Italy, knowing the distribution of their scientific production in the subject categories, can identify the corresponding SDS and select relevant statistic parameters as benchmark for comparative evaluation of their publication rates.

The statistics from the current analyses very clearly demonstrate the heterogeneity of publication rates even in the fields belonging to a single discipline. They help recognize the extent of distortions that occur when comparing the research performance of scientists from different fields, and could then serve as normalization factors to reduce such distortions when comparing the research performance of individuals, groups, or entire research organizations.

In future extensions of this work we could envisage a longitudinal analysis to assess the trends in publication intensity by field. We also know that publication rates of full, associate and assistant professors are different (Abramo, D'Angelo, & Di Costa, 2011). Gender differences in productivity have been demonstrated as well (Abramo, D'Angelo, & Caprasecca, 2009; Leahey, 2006; Fox, 2005; Pripić, 2002; Long, 1992). Because the composition of research staff by academic rank and gender varies across fields, a further extension of the analysis may then entail examining the differing publication intensity across fields by academic rank and gender.

References

Abramo, G., D'Angelo, C.A., & Caprasecca, A. (2009). Gender differences in research productivity: a bibliometric analysis of the Italian academic system. *Scientometrics*, *79*(3), 517-539.

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., & 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., & Rosati, F. (2013a). The importance of accounting for the number of co-authors and their order when assessing research performance at the individual level in the life sciences. *Journal of Informetrics, 7*(1), 198–208.

Abramo, G., D'Angelo, C.A., & Rosati, F. (2013b). Measuring institutional research productivity for the life sciences: the importance of accounting for the order of authors in the byline. *Scientometrics*, *97*(3), 779-795.

Butler, L. (2007). Assessing university research: A plea for a balanced approach. *Science and Public Policy*, *34*(8), 565-574.

D'Angelo, C.A., Giuffrida, C., & Abramo, G. (2011). A heuristic approach to author name disambiguation in large-scale bibliometric databases. *Journal of the American Society for Information Science and Technology*, *62*(2), 257–269.

Egghe, L. (2005). Relations between the continuous and the discrete Lotka power function. *Journal of the American Society for Information Science and Technology*, *56*(7), 664–668.

Fox, M.F. (2005). Gender, family characteristics, and publication productivity among scientists, *Social Studies of Science, 35*(1), 131–150.

Garfield, E. (1979). Is citation analysis a legitimate evaluation tool? *Scientometrics, 1*(4), 359-375.

Kyvik, S. (1989). Productivity differences, fields of learning, and Lotka's law. *Scientometrics*, *15*(3-4), 205-214.

Leahey, E. (2006), Gender differences in productivity: research specialization as a missing link, *Gender and Society, 20*(6), 754-780.

Long, J.S. (1992), Measure of sex differences in scientific productivity, *Social Forces, 71*(1), 159-178.

Lotka, A.J. (1926). The frequency distribution of scientific productivity. *Journal of the Washington Academy of Sciences*, 16 (12), 317–324.

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

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.

Pontille, D. (2004). La Signature Scientifique: Une Sociologie Pragmatique de l'Attribution. Paris: CNRS Éditions.

Pripič, K. (2002). Gender and productivity differentials in science, *Scientometrics, 55*(1), 27-58.

RIN (Research Information Network) (2009). *Communicating Knowledge: How and Why Researchers Publish and Disseminate Their Findings*. London, UK: RIN. Retrieved April 8, 2015 from www.jisc.ac.uk/publications/research/2009/communicatingknowledgereport.aspx.

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., Van Leeuwen, T.N., Van Raan, A.F.J., Visser, M.S., & 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), 2419-2432.

A Technology Foresight Model: Used for Foreseeing Impelling Technology in Life Science

Yunwei Chen*, Yong Deng, Fang Chen, Chenjun Ding, Ying Zheng and Shu Fang

* *chenyw@clas.ac.cn*
Chengdu Library of the Chinese Academy of Sciences, Chengdu, 610041 (China)

Abstract
This paper constructs an Impelling Technology Foresight Model (ITFM) for foreseeing impelling technology in the field of life science, which is a comprehensive model consisting of four class indicators: international scientific environment, evolving of papers and patents, collaboration features of patent assignees' collaboration networks, and impacts. A case study was carried out in the field of life science. Recombinant DNA (RbDNA) and Monoclonal Antibody (mAb) were selected as impelling technologies to carry out the case study. ELISA Diagnosis (ELISA) and Fermentation Technology (FT) were defined as non-impelling technologies to be control group. Results revealed that impelling technologies have higher evolving rates from the stage of growth to maturity. Significant policies or programs usually boost the rapid progress of impelling technologies. Impelling technologies have much higher impact than non-impelling ones. Collaboration behaviour is much more broad and general for impelling technologies. To our knowledge, this is the first study carried out to date to foreseeing impelling technologies at this way.

Conference Topic
Methods and techniques

Introduction
Technology has made enormous contributions to modern society and many future social developments can be realized only through better technical developments and better management (Compton, 1939). Nevertheless, not all technical progress makes substantial contributions to social development. Only a few techniques brought revolutionary change to the human society, such as Transistor Technology and Recombinant DNA technique, which belong to the field of information technology and biotechnology, respectively. Information technology and biotechnology are also regarded as dominant technologies and will essentially impel the social development in the 21st century (Das, 2001).

Thus, it is an attractive topic all the time for scientists from many scientific fields to foresee what kind of technologies can become such impelling technologies, especially in the field of biotechnology. Impelling technology is defined in this paper as technologies that can bolster, lead and push the scientific development and technology progress in given fields, and that can drive the industry fast development and breed emerging industry. Transistor Technology and Recombinant DNA technique are just such technologies. However, people desire to know which technologies can become impelling technologies in the near future, especially for new technologies. For example, synthetic biology, which uses unnatural molecules to reproduce emergent behaviours from natural biology with the goal of creating artificial life (Benner & Sismour, 2005), is recognized as a powerful technique that can produce re-engineered organisms that will change our lives over the coming years, leading to cheaper drugs, green fuel and targeted therapies for diseases. The de novo engineering of genetic circuits, biological modules and synthetic pathways is beginning to address these crucial problems (Khalil & Collins, 2010). If that is true, synthetic biology could be regarded as an impelling technology. However, except for synthetic biology, there are still a large number of techniques emerging in the field of biology. Which can become impelling technology in the near future? Foresight analysis provides the idea of solutions.

Technology foresight, like technology forecasting, is the generation of reasoned statements about the future, the interpretation of such statements in terms of informed action, and the collective learning processes that are involved in responding to challenges of the future (Salo & Cuhls, 2003). Amanatidou (2014) pointed out that the major impacts of foresight belong to knowledge, network creation and promoting public engagement in policy-making. The scope of technology foresight comprises not only technologies and their applications but also public policies and societal challenges (Salo & Cuhls, 2003). UNIDO defined technology foresight as the most upstream element of the technology development process. It provides inputs for the formulation of technology policies and strategies that guide the development of the technological infrastructure. In addition, technology foresight provides support to innovation, and incentives and assistance to enterprises in the domain of technology management and technology transfer, leading to enhanced competitiveness and growth (UNIDO, 2014).

Indeed, similar forms of foresight technology also include technology intelligence, technology forecasting, road mapping and assessment (Firat, 2008). Many of these forms use similar tools and get similar results. Particularly forecasting and foresight are often confused in practice. According to the interpretation from the Technology Futures Analysis Methods Working Group 1 (TFAMWG), all these similar methods could be used in technology futures analysis (TFA). Technology foresight is used to analyse the effecting development strategy, often involving participatory mechanisms. Technology forecasting is to anticipate the direction and pace of changes. But there is a general tendency that forecasting usually focuses on specific technologies. Foresight studies usually bring together people with different expertise and interests, and use instruments and procedures that allow participants to simultaneously adopt a micro view of their own disciplines and a systems view of overriding or shared objectives (Firat, 2008). Some foresight related studies are introduced below and their findings contributed partly to the theoretical and technical basis of this study.

Based on the below related works analysis, we found that although many techniques have been used to answer many kinds of questions, impelling technology foresight works were lacking, especially by the method of model construction. Therefore, this study advanced the existing works by constructing an ITFM model to carry out impelling technology foresight analysis. ITFM model can be used for impelling technology foresight. To our knowledge, none of the existing studies has done such work as ever. The significance of this work is that if an impelling technology could be known before it becomes impelling technology or at the earlier stage of its life cycle, that would be very valuable for many kinds of scientists, policy makers and stakeholders to deal with it.

Related works

The term “Technology Foresight” was introduced by Irvine and Martin and took off in the 1990s as European, and then other countries (Miles, 2010). Until now, a lot of studies have been carried out to do such analysis in recent years, which could be divided into four aspects: function, subject areas of use, features of products and results, and techniques. Related works are discussed below.

Function

The focuses of technology foresight studies have been often motivated by the desire to shape S&T policies and analyse the challenges of education, services, health, and environment, etc. (Salo & Cuhls, 2003). For example, Carlson (2004) discussed the using of technology foresight to create business value. Sanz-Menendez (2001) made technology foresight as a useful tool for policy making. Havas (2010) analysed the impact of foresight on innovation policy-making. Weigand et al. (2014) studied collaborative foresight method to complement long-horizon strategic planning.

Subject areas of use

Based on the fields of science and technology, Linstone (2011) discussed the unique impacts of technology foresight on nanotechnology, biotechnology and materials science. Weinberger, Jorissen and Schippl (2012) carried out a study about technology foresight analysis in the field of environmental technologies with the purpose of supporting the process of identifying and recommending options for the prioritisation of future research funding. Furthermore, foresight has also been used in the field of education studies (Goldbeck & Waters, 2014; King, 2014), drugs discovery (Lintonen et al., 2014).

Features of products and results

From the aspect of products and results of foresight, the works of technology foresight usually have the following products: Strategic advice or guidance, particular technologies or their consequences, price or trends of markets, and production. For example, Cook, Inayatullah and Burgman (2014) concluded that foresight could play a more significant role in environmental decisions by the following ways: monitoring existing problems, highlighting emerging threats, identifying promising new opportunities, testing the resilience of policies, and defining a research agenda. Markus and Mentzer (2014) discussed the future consequences of ICT. Weinberger, Jorissen and Schippl (2012) used foresight methods to support the process of identifying and recommending options for the prioritisation of future research funding among the wide range of environmental technologies available that can contribute to progress in the field of environment.

Techniques

At the angle of techniques used for foresight, many kinds of methods have been used to carry out technology foresight analysis. One typical technique is bibliometric methods. Van Raan (1996) overviewed the potentials and limitations of bibliometric methods for the assessment of strengths and weaknesses in research performance, and for monitoring scientific developments. The study suggested that research performance assessment is based on advanced analysis of publication and citation data. While for monitoring scientific developments, bibliometric mapping techniques are essential. Actually, mapping has been widely used for technology foresight. For example, Yoon, Lee and Lee (2010) developed a keyword-based knowledge map to use to establish a policy to support promising R&D areas and devise a long-term research plan. Another typical method is modelling and system. For instance, Shiue and Lin (2011) developed a foresight MASA model for future technology evaluation in electric vehicle industry, which integrated the concept of vision, linking analysis planning, Markov chain, and Scenario analysis (SA). Chen (2012) proposed a structural variation model for answering what kinds of information may serve as early signs of potentially valuable ideas. Peer review and Delphi have also been used in foresight as in forecasting. For example, Lintonen et al. (2014) had done a drugs foresight analysis in 2020 through the method of Delphi expert panel study. Forster & Gracht (2014) had also assessed Delphi panel composition for strategic foresight based on company-internal and external participants.

Model of Impelling Technology Foresight Model (ITFM)

Definition and Hypothesis

As is stated above, impelling technologies are such technologies that could bolster, lead and push the scientific development and technology progress and drive the existing industry fast develop and bread emerging industry in given fields. However, this definition explains only the functional feature reflecting the results generated by impelling technologies, and lacks the

description of its inherent features, especially the features at the early stage of technology lifetime, which are much more important to foresee whether a technology at the early stage could become impelling technologies. Therefore, the inherent features of impelling technologies especially the features at the early stage could be used as indicators for reflecting impelling technologies. Thus, some hypothesises had been proposed as the theoretical base for constructing an Impelling Technology Foresight Model (ITFM) for foreseeing impelling technologies, particularly in the field of life science.

Hypothesis 1. Viewed by the concept of technology life cycle, technologies' development process can be divided into four stages (Little, 1981) of emerging, growth, maturity and saturation. Impelling technologies grow rapidly to the stage of maturity after short growth stage. Impelling technologies seldom show signs of turning to saturation stage for their competitive impact could remain much longer than non-impelling technologies. In order to evaluate the current stages of a technology, patents have been widespread used to do such analysis. For example, Patent analysis was applied by Zhou et al. (2014) to monitor the developmental stage of a particular New and Emerging Science & Technologies, dye-sensitized solar cells (DSSCs), and traced its potential evolutionary pathways. Some other related works have high impacts include Haupt, Kloyer & Lange (2007), Trappey & Wu (2011), Jarvenpaa, Makinen & Seppanen (2011), etc. This paper uses patent data to disclose the different/given features at the different stages of impelling technologies.

Hypothesis 2. During the development process of an impelling technology, pushing policies or programs usually would like to be attracted to boost the progress of impelling technology. For example, Human Genome Project has been the first major foray of the biological and medical research communities and it boosted the development of an array of new technologies (Collins, Morgan & Patrinos, 2003), among which Recombinant DNA technique have achieved considerable development and have also been generally recognized as an impelling technology in the field of life science.

Hypothesis 3. Impelling technologies have higher level of collaboration, especially in patent assignees' collaboration. A lot of studies have shown that there is a positive correlation between collaboration and better production of science. For instance, Guimerà, et al. (2005) pointed out that collaboration could spur creativity, solving old problems and inspiring fresh thinking. In the field of scientific researches, Whitfield (2008) pointed out that there is a picture of science's increasingly collaborative nature and which determine a team's success. Wuchty, Jones and Uzzi (2007) found that there's something about between-school collaboration that's associated with the production of better science. Kato & Ando (2013) found a positive correlation between their research performance and degree of internationalization.

Hypothesis 4. Impelling technologies have higher level of impacts. Citation-based analysis is the most frequently used method to carry impact analysis. The original use of citation for evaluation is Journal Citation Reports from Thomson Reuters to evaluate journals impact factors. Garfield (1979) pointed out that citation analysis could introduce a useful measure of objectivity into the evaluation process at relatively low financial cost. Numerous approaches have been devised to assess future technological impacts based on patent citation information with the core purpose of identifying the current technologies that will drive technological changes over the coming few years (Lee et al., 2012). There are also some network-based method were used to do technology impact analysis. For example, Ko et al. (2014) presented a combined approach for constructing a technology impact network basing on patent co-classification and identifying the impact and intermediating capability of technology areas from the perspective of a national technology system. This paper uses paper citations to compare the difference of impacts between impelling technologies and non-impelling technologies.

ITFM frame

A few factors from four aspects were introduced to validate the above hypothesis.

Technology life cycle - Evolving of patents and paper were introduced to disclose the evolving features of impelling technologies during the four stages of emerging, growth, maturity and saturation.

International environment - The ITFM model took only policy, plan or program as indicators to reflect the international scientific environment although the related factors are more.

Collaboration - The following network statistics of patent assignees collaboration networks were used to represent the collaboration features of impelling technology.

- Ratio of isolates, which have no collaborators in the assignees collaboration networks G. Counted as n (isolates)/n.
- Ratio of nodes in the largest cluster, counted as n (largest cluster)/n.
- Ratio of clusters compare to nodes, counted as #clusters/n.
- Average degree, let N(i) be 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 $\eta(G)=\Sigma_{i \square N}\eta(i)/n$.
- Diameter, which is measured by shortest-path length, has been used to estimate the stage of development through documentation data (Chen, Borner & Fang, 2013, Bettencourt, Kaiser & Kaur, 2009) or patent data (Chen & Fang, 2014). There is a theory that collaboration graph that densify with constant or decreasing diameters. All these studies have 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 (Bettencourt, Kaiser & Kaur, 2009). The assignees collaboration network diameters seem to stabilize at about 12 when a technology come into the stage of maturity (Chen & Fang, 2014).

Note that n is the total number of nodes in the network.

Impact - Two factors of times cited per paper and times cited per patent were used for expressing the technology impacts.

The ITFM frame is listed in Table 1, which is the origin of the following case study.

Table 1. Factors contributing to the ITFM.

Factors		*For validating hypothesis (purpose)*
Technology life cycle	evolving of papers	hypothesis 1
	evolving of patents	
International scientific environment	policy	hypothesis 2
	plan or program	
Collaboration-patent assignees collaboration networks	ratio of isolates	hypothesis 3
	ratio of nodes in the largest cluster	
	ratio of clusters compare to nodes	
	average degree	
	diameter	
Impact	times cited per paper	hypothesis 4
	times cited per patents	

Data and methods

According to the opinions of thirty experts in the field of life science through email consultation, Recombinant DNA (RbDNA) and Monoclonal Antibody (mAb) were selected as impelling technologies to carry out case study. ELISA Diagnosis (ELISA) and Fermentation Technology (FT) were defined as non-impelling technologies to be control group.

Publications in Web of Science^{TM} from 1960s to 2012 (publication year) and US patents in Derwent Innovations Index^{SM} from 1970s to 2012 (basic patent year, defined by DII based on the earliest year of all the publication dates of all members of a patent family) were chosen as quantitative data of case study. Data was acquired from the Web of Science in May 2013. Thomson Data Analyzer (TDA) and Science of Science (Sci^2) Tool (http://cns.iu.indiana.edu) were used to extract the statistic and network information.

Search terms to retrieval papers and patents are listed in Table 2.

Table 2. Search terms used for this study.

	Papers	*Patents*
RbDNA	TS=("DNA recombination" or "recombinant DNA" or "DNA cloning" or "molecular cloning" or "gene cloning")	IPCs: from C12N-015/09 to C12N-015/90
mAb	TS=(("monoclon* antibod*") OR (monoclon* same antibod*))	IP=C12P-021/08
FT	TS=ferment*	IP=(C12C-011/* OR C12G* OR C12P* OR C12J*) AND TS=ferment*
ELISA	TS=elisa, removed the papers in WC class of Spectroscopy, Optics, Physics Condensed Matter, Nuclear Science Technology, Behavioral Sciences, Astronomy Astrophysics and Microscopy.	TS=Elisa

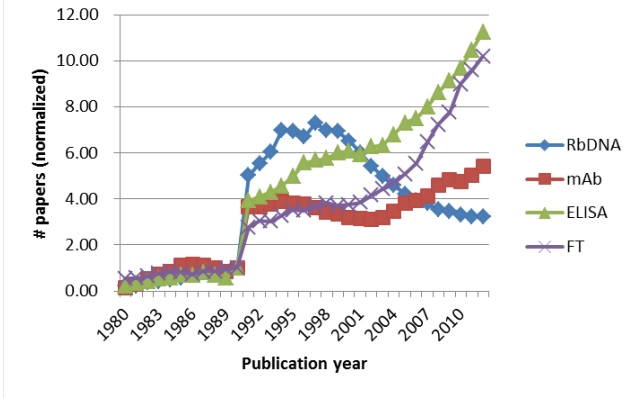
Results and Analysis

Evolving of papers and patents

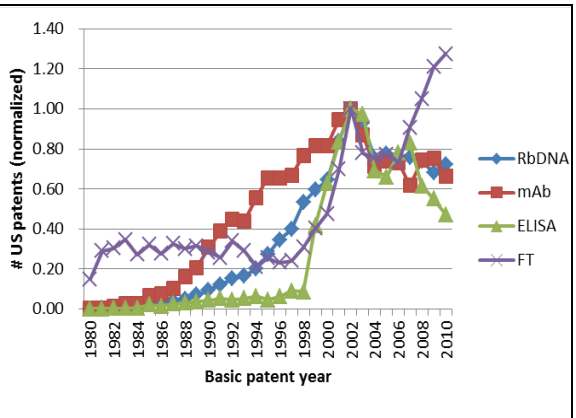
Papers and patents are two external indicators for reflecting the evolving of technologies. The output of papers and patents of the two impelling technologies and two non-impelling technologies were normalized to 1 by their numbers of papers in 1990 and numbers of patents in 2002 separately. The reason of choosing 1990 was that the year 1990 was a jumping-off year, after when the number of papers jumped at least more than three times in 1991. The reason of choosing 2002 was that the year 2002 was a dividing crest, which year had the maximum number of patents, except for FT. Fig. 1 illustrates that the number of papers of both the two impelling technologies stabled at a certain range after three or four years development following the jumping-off from 1990 to 1991. The patents trends show that the number of patents of impelling technologies stabled at a certain level after two years of the patent outputs peak. However, both the papers trends and patent trends of non-impelling technologies had no stable signal no matter which way they go, increase or decrease constantly.

In order to compare the features of impelling technologies at different stages of life cycle, time were sliced into four sections, -1986 (emerging stage), 1987-1993 (growth stage), 1994- (maturity and saturation stages). This division mainly depended on the evolving histories of the two impelling technologies. Although it was not adaptive for on-impelling technologies it

had also been used for distinguishing non-impelling technologies' life cycles with the purpose of comparison.



papers-normalized by the number of papers in 1990



patents-normalized by the number of patents in 2002

Figure 1.Growth of papers and US patents.

International scientific environment

Through watching the histories of the two impelling technologies, we found that Human Genome Project, the first major foray of the biological and medical research communities launched in 1990, boosted the two impelling technologies fast into maturity stage, which could be reflected by the jump of the number of papers. Nevertheless, although the two non-impelling technologies, ELISA and FT, had also been boosted by the Human genome Project, these two technologies had not entered into maturity stage throughout. Actually, beside for the Human Genome Project, there were still more crucial policies had been drawn and put into effect. For example, USA had announced the first Recombinant DNA research Guidelines for normalizing such researches. Even till now, government still made positive policies to maintain the driving functions of impelling technologies. For instance, US Federal Court ruled that synthetic DNA could be patented, which might become a new pushing for the development of RbDNA.

In the aspect of industry, at the stage of growth there were one or a few professional companies born and the number of companies rose sharply at the stage of development and the early maturity stage. For instance, benefited from the development of RbDNA, the first biotechnology company Genentech had been established in 1976. When an impelling technology is mature, the relevant industry would expand rapidly. For example, mAb had brought a rapid growth market of 26 billion USD in 2006 while it was only 4 billion in 2002.

Patent assignees collaboration networks

Figure 2 and Figure 3 illustrate the network features of patent assignees collaboration networks. It is clearly showed in Figure 2c that as time gone on, the ratio of isolates (assignees have no collaborators) decreased year by year and seemed to stabled at a certain level. However, the ratios of isolates of the impelling technologies were much lower all along than that of the non-impelling technologies. The values of the latter were more than twice of the former. The gap was enlarged to more than three times at the stage of development. As a result of the reduction of isolates, the clusters increased and there were many a big cluster became bigger and bigger. It has to be noted that an isolate was also regarded as a cluster. Therefore, a network with high level of collaborative behaviours must has less clusters because of much more isolates and small clusters tend to merge to bigger clusters. Thus the

excellent performance of collaboration leads to generate a super big cluster and less ration of clusters (see Fig. 2a). Figure 2b shows that the biggest cluster of impelling technologies gathered about more than half of the total number of assignees particularly after the stage of development, which was much higher than that of the non-impelling technologies.

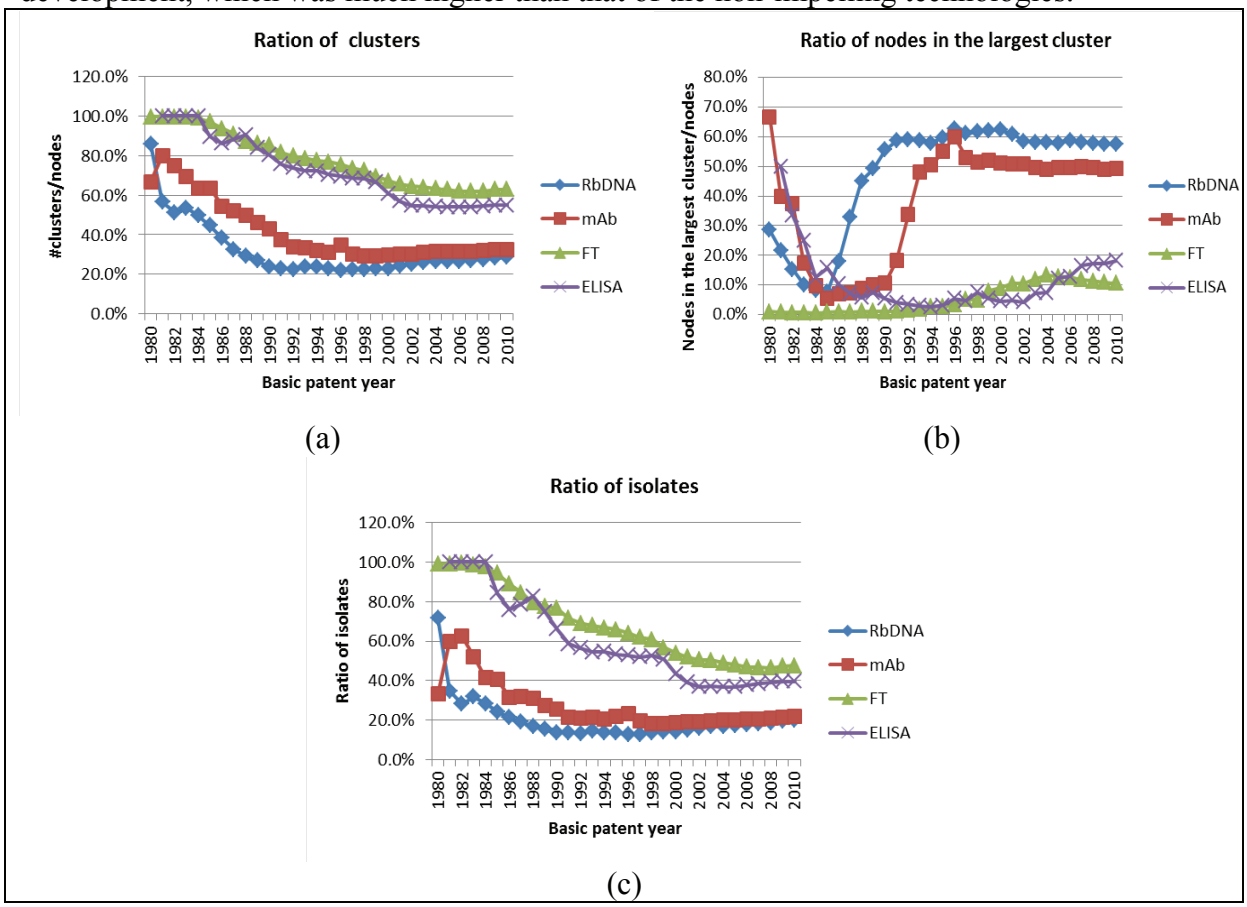


Figure 2. Network features of US patents' assignees' collaboration.

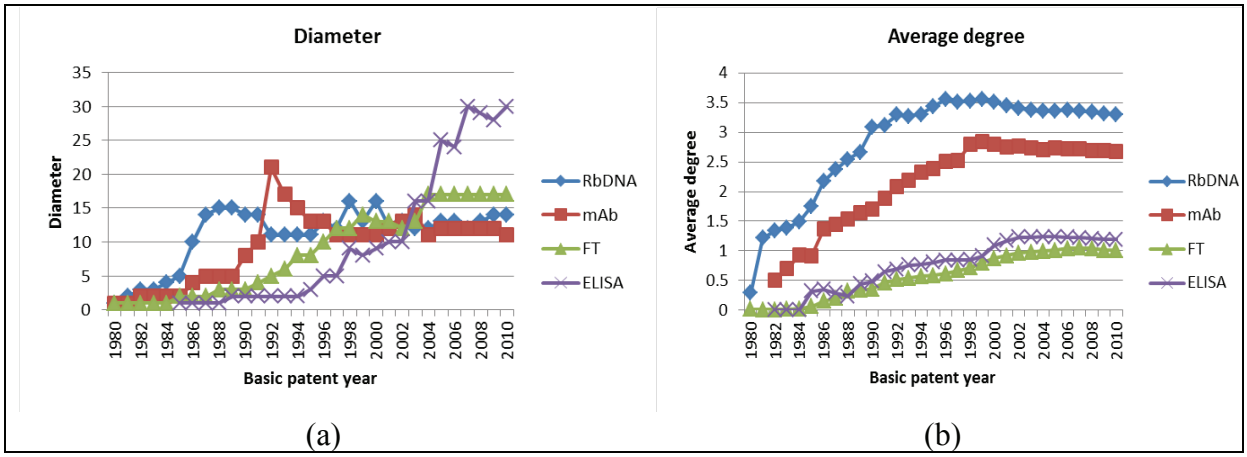


Figure 3. Diameters and average degrees of assignees collaboration network.

Benefited from the good network performance, the impelling technologies had higher average degree all the time. It was about three times higher than that of the non-impelling technologies at the stage of maturity, ten and four times during the period of growth and development respectively.

Impact

The average times cited of papers and patents of the two impelling technologies and two non-impelling technologies were illustrated in Figure 4.

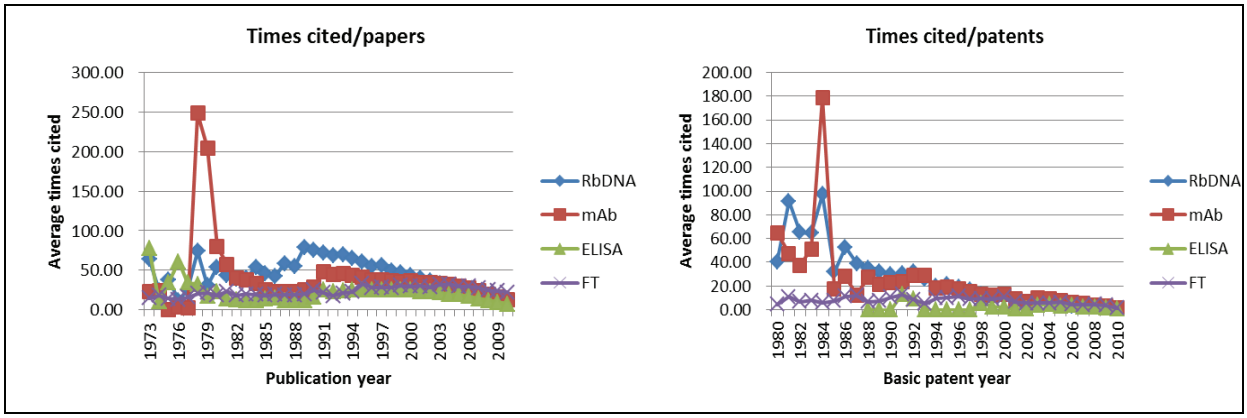


Figure 4. Average times cited of papers and US patents.

The results showed that the average times cited of papers of impelling technologies was two times higher than non-impelling technologies during the whole period of this analysis. The value of impelling technologies was 30 and 50-80 compared to 18 and 20 of non-impelling technologies at the stages of growth (before 1986) and development (1987-1993). For patents, the average times cited of impelling technologies and non-impelling technologies were 66 and 7 in stage of growth, 24 and 9 in stage of development correspondingly. However, the advantages of impelling technologies were eroded as time goes on in the stage of maturity.

Quantitative ITFM

Through the above case study we might conclude some unique features of impelling technologies in the field of life science.

First, impelling technologies had higher rates of evolution from the stage of growth to maturity, which could be illustrated particularly by the papers evolving patterns. When it comes to the technologies of RbDNA and mAb, it took only about one year that both of the two impelling technologies had finished their transform. At the same time, impelling technologies represented distinct feature at the stage of maturity. Nevertheless, the two non-impelling technologies represented no obvious such transformation. It seemed like that both the two non-impelling technologies were still at the stage of growth. However, the fermentation technology had a much longer history than both the two impelling technologies. The reason of it represented such evolutionary feature might just due to the position as a non-impelling technology, which contributes more and more to the society development, but always is a applied technology and will not play more impelling functions.

Second, significant policies or programs boosted the rapid progress of impelling technologies. Although non-impelling technologies had also been pushed by specific policies or plans, the range was lower than that of impelling technologies. When the impelling technologies switched into maturity stage, they usually drove the explosive increase of industry.

Third, impelling technologies had much higher impact than non-impelling technologies, which could be reflected by the times cited per paper/patent. The value of times cited per paper/patent of impelling technologies was two to three times higher than non-impelling technologies. It was highlighted during the process of involving from the stage of development to maturity. In the case of life science, for papers, the value of impelling technologies was 50-80 compared to 20 of non-impelling technologies, for patents, the values were 24 and 9 correspondingly.

Last, collaboration behaviour measured by the collaborations of patent assignees was much more broad and general for impelling technologies. Assignees collaboration networks of impelling technologies had fewer isolates, and there were only about 20% assignees were isolates at the stages of development and maturity. Much more assignees had collaborated with others and become much bigger clusters with a result of the number of clusters decreased. The biggest cluster (principal component) gathered a large number of assignees that took up more than half of the total number of all nodes in the networks at the stage of maturity. As a result, the average degree of impelling technologies reached to 3 which were three times to that of non-impelling technologies at the stage of maturity. The diameters of impelling technologies stabilized at 12 at the stage of maturity. Non-impelling technologies had no such features of stable diameters.

The results indicate that hypothesises listed above were answered by the case study. Based on the results of the comparison of impelling technologies and non-impelling technologies in the field of life science, a quantitative model is induced in table 3. The model can be used for foreseeing any new impelling technologies that have just born or at different stages, especially at the stages of development and maturity.

Table 3 Quantitative ITFM.

	Indicators	***Features***		
		Growth (- 1986)	***Development (1987-1993)***	***maturity (1994-)***
International scientific environment	**Policies, plans &projects**	New incentive, convenient policies enacted	Pushed significantly by major project	Still focus of policies, plans & projects
	Industry	Start-up companies	Number of companies would rise sharply	Industry expand rapidly
Evolving of papers and patents	**Papers evolution**	/	Evolved into maturity stage in few years	Stable (no sign of stable)
	Patents evolution	Steady increase	Steady increase	Stable(no sign of stable)
Collaboration-Features of patent assignees collaboration networks	**Ratio of isolates**	40% (*95%*)	20% (*70%*)	20% (*50%*)
	Nodes in the largest cluster/nodes	20% (*10%*)	35% (*3%*)	55% (*10%*)
	#clusters/#nodes	60% (*97%*)	35% (*80%*)	30% (*65%*)
	Average degree	1 (*0.1*)	2 (*0.5*)	3 (*1*)
	diameters	3 (*1*)	12 (*3*)	Stable at 11-14 (no sign of stable)
Impacts	**average times cited of papers**	30 (*18*)	50～80 (*20*)	Decreased yearly
	average times cited of patents	66 (*7*)	24 (*9*)	No difference between impelling and no-impelling technologies

Notes. The values of non-impelling technologies were listed in brackets.

Discussions

This paper defines impelling technologies and constructs an ITFM model for foreseeing technologies that have potential to become impelling technologies. There is no doubting that this is an attractive topic all the time for many kinds of scientists, policy makers and stakeholders. The theoretical basis of this study is the positive correlation between the four hypothesises and the performance of an impelling technology. Four classes of indicators were introduced into the ITFM model and demonstrated on two impelling technologies and two contrasted non-impelling technologies in the field of life science. Indeed, this work is the first study about impelling technologies foresight and got some valuable results which could be

used for many new technologies foresight, such as synthetic biology. Such application study would be carried out in the near future.

Nevertheless, there are still some shortages of this study. First, the ITFM model can be used only for evaluating existed technologies and not for future technologies that have not born yet. Indeed, this topic is also interesting and important. Second, the values in the ITFM were concluded from the four technologies from life science, which might volatile when used in other fields. Actually, different impelling technologies even in the field of life science might get different values. Therefore, the values in ITFM model are referenced values. The relative performance of impelling technologies is more important when the model is used for evaluating other technologies. Third, impelling technologies foresight is a complex question, which is hard to be identified easily through one or two models or methods. There must be many other indicators that could reflect the unique features of impelling technologies. Therefore, this work is just a beginning of such efforts for foreseeing impelling technologies.

Acknowledgements

We would like to thank many fellows in the field of life science from CAS, who gave a lot of advises in choosing the technologies used for case study. The paper did benefit greatly from detailed comments by an anonymous reviewer. This work is funded by the Documentation and Information Special Project of Chinese Academy of Sciences (2013). This work is funded in part by the National High Technology Research and Development Program of China (863 Program) under grant no. 2014AA021503. This work is supported in part by the West Light Foundation of the Chinese Academy of Sciences, China under grant no. [2013]165(3-6). This work is funded in part by the Main Direction Program of Knowledge Innovation of Chinese Academy of Sciences (KSCX2-EW-G-9).

References

Amanatidou, E. (2014). Beyond the veil - The real value of Foresight. *Technological Forecasting and Social Change*, *87*, 274-291.

Benner, S. A. & Sismour, A. M. (2005). Synthetic biology. *Nature Reviews Genetics*, *6*(7), 533-543.

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.

Carlson, L. W. (2004). Using technology foresight to create business value. *Research-Technology Management*, *47*(5), 51-60.

Chen, C. M. (2012). Predictive effects of structural variation on citation counts. *Journal of the American Society for Information Science and Technology*, *63*(3), 431-449.

Chen, Y. W., Börner, K. & Fang, S. (2013). Evolving collaboration networks in Scientometrics in 1978-2010: a micro-macro analysis. *Scientometrics*, *95*(3), 1051-1070.

Chen, Y. W. & Fang, S. (2014). Mapping the evolving patterns of patent assignees' collaboration networks and identifying the collaboration potential. *Scientometrics*, *101*(2), 1215-1231.

Collins, F. S., Morgan, M. & Patrinos, A. (2003). The human genome project: Lessons from large-scale biology. *Science*, *300*(5617), 286-290.

Compton, K. T. (1939). Technical progress and social development. *Electrical Engineering*, *58*(1), 12-15.

Cook, C. N., Inayatullah, S., Burgman, M. A., et al. (2014). Strategic foresight: how planning for the unpredictable can improve environmental decision-making. *Trends in Ecology & Evolution*, *29*(9), 531-541.

Das, M. R. (2001). Biotechnology in the 21(st) Century. *Defence Science Journal*, *51*(4), 327-332.

Firat, A. K. (2008). Technological Forecasting – A Review. Working Paper CISL# 2008-15.

Forster B. & Gracht H. (2014). Assessing Delphi panel composition for strategic foresight - A comparison of panels based on company-internal and external participants. *Technological Forecasting and Social Change*, *84*, 215-229.

Garfield, E. (1979). Is citation analysis a legitimate evaluation tool. *Scientometrics*, *1*(4), 359-375.

Goldbeck, W. & Waters, L. H. (2014). Foresight education: When students meet the future(s). *Futurist*, *48*(5), 30.

Guimerà, R., Uzzi, B., Spira, J. & Amaral, L. A. N. (2005). Team assembly mechanisms determine collaboration network structure and team performance. *Science*. *308*, 697–702.

Haupt, R., Kloyer, M. & Lange, M. (2007). Patent indicators for the technology life cycle development. *Research Policy*, *36*(3), 387-398.

Havas, A., Schartinger, D. & Weber, M. (2010). The impact of foresight on innovation policy-making: recent experiences and future perspectives. *Research Evaluation*, *19*(2), 91-104.

Jarvenpaa, H. M., Makinen, S. J. & Seppanen, M. (2011). Patent and publishing activity sequence over a technology's life cycle. *Technological Forecasting and Social Change*, *78*(2), 283-293.

Kato, M. & Ando, A. (2013). The relationship between research performance and international collaboration in chemistry. *Scientometrics*, *97*(3), 535–553.

Khalil, A. S. & Collins, J. J. (2010). Synthetic biology: applications come of age. *Nature Reviews Genetics*, *11*(5), 367-379.

King, K. (2014). Foresight in middle school: Teaching the future for the future. *Futurist*, *48*(5), 41-42.

Ko, S. S., Ko, N., Kim, D., et al. (2014). Analyzing technology impact networks for R&D planning using patents: combined application of network approaches. *Scientometrics*, *101*(1), 917-936.

Lee, C., Cho, Y., Seol, H., et al. (2012). A stochastic patent citation analysis approach to assessing future technological impacts. *Technological Forecasting and Social Change*, *79*(1), 16-29.

Lintonen, T., Konu, A., Ronka, S., et al. (2014). Drugs foresight 2020: a Delphi expert panel study. *Substance Abuse Treatment Prevention and Policy*, 9.

Little, A. D. (1981). *The Strategic Management of Technology*. Cambridge, Mass.

Mansfield, E. (1961). Technical change and the rate of imitation. *Econometrica*, *29*(4), 741-766.

Markus, M. L. & Mentzer, K. (2014). Foresight for a responsible future with ICT. *Information Systems Frontiers*, *16*(3), 353-368.

Miles, I. (2010). The development of technology foresight: A review. *Technological Forecasting and Social Change*, *77*(9), 1448-1456.

Salo, A. & Cuhls, K. (2003). Technology foresight—past and future. *Journal of Forecasting*, *22*(2-3), 79-82.

Sanz-Menendez, L., Cabello, C. & Garcia, C. E. (2001). Understanding technology foresight: the relevance of its S & T policy context. *International Journal of Technology Management*, *21*(7-8), 661-679.

Shiue, Y. C. & Lin, C. Y. (2011). Developing a new foresight model for future technology evaluation in electric vehicle industry. *Journal of Testing and Evaluation*, *39*(2), 119-125.

Trappey, C. V., Wu, H. Y., Taghaboni-Dutta, F., et al. (2011). Using patent data for technology forecasting: China RFID patent analysis. *Advanced Engineering Informatics*, *25*(1), 53-64.

UNIDO. (2014). *Technology Foresight*. Retrieved May 10, 2014 from: http://www.unido.org/foresight.html.

Van Raan, A. F. J. (1996). Advanced bibliometric methods as quantitative core of peer review based evaluation and foresight exercises. *Scientometrics*, *36*(3): 397-420.

Weigand, K., Flanagan, T., Dye, K., & Jones, P. (2014). Collaborative foresight: Complementing long-horizon strategic planning. *Technological Forecasting and Social Change*, *85*, 134-152.

Weinberger, N., Jorissen, J. & Schippl, J. (2012). Foresight on environmental technologies: options for the prioritisation of future research funding - lessons learned from the project "Roadmap Environmental Technologies 2020+". *Journal of Cleaner Production*, *27*, 32-41.

Whitfield, J. (2008). Collaboration: Group theory. *Nature*. *455*, 720-723.

Wuchty, S., Jones, B. F. & Uzzi, B. (2007). The Increasing dominance of teams in production of knowledge. *Science*. *316*, 1036–1039.

Yoon, B., Lee, S. and Lee, G. (2010). Development and application of a keyword-based knowledge map for effective R&D planning. *Scientometrics*, *85*(3), 803-820.

Zhang, J. X., Zhang, H. S., de Pablos, P. O., et al. (2014). Challenges and foresights of global virtual worlds markets. *Journal of Global Information Technology Management*, *17*(2), 69-73.

Zhou, X., Zhang, Y., Porter, A. L., et al. (2014). A patent analysis method to trace technology evolutionary pathways. *Scientometrics*, *100*(3), 705-721.

Lung Cancer Researchers, 2008-2013: Their Sex and Ethnicity

Grant Lewison[1], Philip Roe[2] and Richard Webber[3]

[1] *grant.lewison@kcl.ac.uk*
King's College London, Guy's Hospital, Great Maze Pond, London SE1 6RT (UK)

[2] *philip@evaluametrics.co.uk*
157 Verulam Road, St Albans, AL3 4DW (UK)

[3] *richardwebber@blueyonder.co.uk*
Kings's College London, Department of Geography, Strand, London WC2R 2LS (UK)

Abstract
This paper describes the process by which almost all authors of papers in the Web of Science (WoS) can be characterised by their sex and ethnicity or national background, based on their names. These are compared with two large databases of surnames and given names to determine to which of some 160 different ethnic groups they are most likely to belong. Since 2008 the authors of WoS papers are tagged with their addresses, and many have their given names if they appear on the paper, so the workforce composition of each country can be determined. Conversely, the current location of members of particular ethnic groups can be found. This will show the extent of a country's "brain drain", if any. Key results are shown for one subject area, and *inter alia* it appears that the majority of researchers of Indian origin who are active in lung cancer research are working in the USA. But East Asians (Chinese, Japanese and Koreans) tend to stay in their country of birth.

Conference Topic
Methods and techniques

Introduction
There is continuing research interest in the sex and ethnic composition of research personnel. A brief survey of the literature in 2013-2014 indicates that there is a widespread interest in the problems faced by female researchers (no fewer than 24 countries were involved in such research, and there were 71 papers in the two years, including several exploring the problems in countries outwith North America and western Europe (*e.g.*, Gonenc et al., 2013; Homma, Motohashi, & Ohtsubo, 2013; Bettachy et al., 2013; Isfandyari-Moghaddam & Hasanzadeh, 2013; Garg & Kumar, 2014). However there is much less interest in the situation of ethnic groups, and that only in the USA (Griffin, Bennett & Harris, 2013; Pololi et al., 2013; Campbell et al., 2013; Hassouneh et al., 2014), with one exception (Johansson & Sliwa, 2014; Sliwa & Johansson, 2014), which concerned foreign women in a UK business school. Attention in the USA is focussed almost entirely on under-represented minorities (African-Americans, Hispanics, and in some cases Native Americans), and hardly at all on the problems that may be encountered by researchers of Asian origins, notably Chinese and Indians, who may have to cope with difficult immigration (Teich, 2014), integration and living experiences when they move to the USA. In fact, as we shall see, they are hardly "under-represented minorities" but rather over-represented compared with their presence in the population. (A fuller survey of the relevant prior literature was given in Roe et al., 2014.)
This paper provides a method whereby the researchers in a given scientific subject area can be characterised by their ethnicity or national background and their sex. This is important for science policy, including the monitoring of the changing roles and positions of women in research and the extent to which a country is welcoming to researchers from abroad and helps them to integrate. It builds on the methods described earlier (e.g., Roe et al., 2014) but now allows all the authors on multi-national papers to be classified, and is applicable to all the countries represented in the subject area. Conversely, it can reveal the location of researchers

of any particular ethnicity or national origin. The methods have been applied to the subject area of lung cancer research, and results for this area are given in some detail, but they can equally be applied to any other research area.

Attention was focussed on 24 leading countries, responsible for the large majority of global lung cancer research output, as shown in Table 1 with their digraph ISO codes. However, some results are also given for others, because the database listed all countries contributing to lung cancer research, and researchers with names characteristic of 90 different countries.

Table 1. List of 24 leading countries in lung cancer research, 2004-13.

Countries	*ISO*	*Countries*	*ISO*	*Countries*	*ISO*	*Countries*	*ISO*
Australia	AU	Denmark	DK	Japan	JP	Sweden	SE
Austria	AT	France	FR	Netherlands	NL	Switzerland	CH
Belgium	BE	Germany	DE	Norway	NO	Taiwan	TW
Brazil	BR	Greece	GR	Poland	PL	Turkey	TR
Canada	CA	India	IN	South Korea	KR	United Kingdom	UK
China (PR of)	CN	Italy	IT	Spain	ES	USA	US

Methodology

The file of lung cancer papers (articles and reviews) was obtained from the Web of Science (WoS) for the six years, 2008-2013, from the intersection of two "filters". One was for cancer, and was based on journal names and title words. These included the names of many individual cancers, genes known to pre-dispose people to an enhanced (or reduced) risk of cancer, and specialist drugs and other treatments such as radiotherapy. The other was for lung disease, and consisted of a number of specialist respiratory journals, such as *Experimental Lung Research, Jornal Brasileiro de Pneumologia, Lung* and *Respiration*, and two title words *lung* and *trachea**. In addition, all the papers in the journals *Lung Cancer* and *Clinical Lung Cancer* were retained, together with papers with *SCLC* or *NSCLC* in their titles. The file contained details of 22,433 papers.

The analysis of the researchers was based on their names, both surnames and given names. The surnames were compared with our listing of 2.6 million family names which is based on records of the majority of the adult population in the following countries: Australia, Brazil, Denmark, Germany, Ireland, Italy, Netherlands, Norway, South Africa, Spain, Sweden, the UK and the USA as well as surname frequency distributions for Austria, Belgium, France, India and Japan. For some countries in Eastern Europe and the Middle East, the files were supplemented by data on the names of scientists from these countries found in the WoS. We were able to classify names into over 160 different ethnicities, nationalities and regions within countries, but in this study the classification was simplified to include own country and eight main groups:

- own country (OWN) – this also included representatives of countries who have been the main sources of immigrants, such as France and the UK in Canada;
- other European country (EUR: Albania, Balkan, Belgium, Bosnia, Britain, Bulgaria, Croatia, Cyprus, Czech, Denmark, Estonia, Finland, France, Germany, Greece, Hungary, Iceland, Ireland, Italy, Latvia, Lithuania, Malta, Montenegro, Netherlands, Nordic, Norway, Poland, Portugal, Romania, Russia, Serbia, Slovakia, Slovenia, Spain, Sweden, Switzerland);
- Latin America (LAT: including Brazil, Guyana and Mexico);
- Levant and Mediterranean (LEV: Algeria, Egypt, Israel, Lebanon, Libya, Morocco, Saudi Arabia, Tunisia, Turkey, Ukraine);

- Africa (AFR: Afrikaaner, Angola, Cameroon, Congo, Eritrea, Ethiopia, Ghana, Ivory Coast, Kenya, Malawi, Mauritius, Nigeria, Sierra Leone, Somalia, Sudan, Uganda);
- South Asia (SAS: Bangladesh, Burma, India, Pakistan, Sri Lanka);
- China (CHI);
- other Asia (ASI: Afghanistan, Armenia, Azerbaijan, Cambodia, Georgia, Iran, Iraq, Japan, Korea, Laos, Malaysia, Mongolia, Nepal, Philippines, Singapore, Thailand, Vietnam);
- other non-European and Oceanic (OCE: Australia, Caribbean, Fiji, Indonesia, New Zealand).

The methodology is more fully described in a recent paper by Roe et al. (2014).

Given names often (but not always) connote the sex of the person, and we have compiled a list of some 0.7 million such names, including some misspellings and phonetic misrepresentations. This has recently been complemented with the given names of all UK doctors on the Medical Register – over 328,000 individuals, many of whom come from other countries. Some given names connote a different sex in different countries – for example, Andrea is female in the UK but male in Italy. A few countries (in the present study, only Poland) have surnames with gender endings and this can also be used to determine the sex of an author.

In (Roe et al., 2014), attention was confined to papers from a single country, but we were now able to identify the names of the authors from each of the countries in a multi-national paper because the WoS lists them with their addresses in the following format:

> *[Scagliotti, Giorgio V.] Univ Torino, Thorac Oncol Unit, Dept Clin & Biol Sci, S Luigi Hosp, I-10043 Turin, Italy; [Germonpre, Paul] Univ Ziekenhuis Antwerpen, Edegem, Belgium; [Planchard, David] CHU Poitiers, Poitiers, France; [Reck, Martin] Krankenhaus Grosshansdorf, Grosshansdorf, Germany; [Lee, Jin Soo] Natl Canc Ctr Korea, Goyang, South Korea; [Biesma, Bonne] Jeroen Bosch Ziekenhuis, Shertogenbosch, Netherlands; [Szczesna, Aleusandra] Mazowieckie Ctr Leczenia Chorob Pluc & Gruzlicy, Otwock, Poland; [Morgan, Bruno] Leicester Royal Infirm, Dept Radiol, Leicester, Leics, England*

although not all the authors have given names that would allow their sex to be determined.

A special macro was written to enable the names of all authors from each of the countries to be listed in appropriate columns of a spreadsheet for each paper. These were then each classified by national group and sex, where available, so that the contributions of each of the national groups and sexes could be determined. However, the main analysis was performed on the long list of 84,533 different names, each of which was associated with a country and had its frequency of occurrence listed. For each of the 24 selected countries, and for the rest of the world (RoW), the composition of the lung cancer research workforce and the contributions (sums of the numbers of papers) from researchers from each ethnic group (or world region) were determined.

However, we found during our analysis that some East Asian names belonging to researchers working in China, Japan or South Korea, had been misclassified as European as they were ambiguous, such as Jung, Lee and Park. It was obvious from the given names of these researchers if they were Orientals or Europeans. Thus Jung, Andreas working in Germany was clearly German, but Jung, Deuk-Kju working in South Korea was Korean. Likewise, Park, Bernard J. working in the USA was considered to be of European origin, but Park, Byung-Joo in Korea was taken as Korean. These were manually corrected, and some other adjustments to ethnicity were made.

It also became apparent that some names with different given names or initials actually referred to the same person. Thus there were only two Aaronsons in our list of researchers,

one was Neil and the other Stuart A. Both could be classed as male. Another Aaronson, S.A. was clearly the same as Aaronson, Stuart A, and so could be counted as male. We were able to sex quite a lot of researchers without given names in this way.

Results

The data on the national origins and on the sex of the lung cancer researchers in the 24 selected countries, plus the Rest of the World, were obtained from a large file that looked like this:

Table 2. Small excerpt from the file listing the names of all lung cancer researchers.

Name	*Country*	*ISO*	*Count*	*Ethnic*	*Sex*	*Region*
Aakre, J.	USA	US	1	NO	M	EUR
Aakre, Jeremiah	China	CN	1	NO	M	EUR
Aakre, Jeremiah A.	USA	US	4	NO	M	EUR
Aamini, Mahnaz	Iran	IR	1	IR	F	ASI
Aapro, M.	Switzerland	CH	1	FI	X	EUR
Aarab-Terrisse, S.	France	FR	1	MA	X	LEV
Aarndal, Steinar	Norway	NO	2	NO	M	EUR
Aaron, Jesse	USA	US	1	UK	M	EUR
Aarons, Y.	Australia	AU	1	ES	F	EUR
Aarons, Yolanda	Australia	AU	1	ES	F	EUR

The top person in this list evidently worked both in China and the USA, and the first and ninth names were sexed by comparison with the row(s) below.

For the analysis by sex, all 24 countries, plus the RoW, have been included in Table 3. The table shows the percentages of names that could be sexed, and the percentage of such names that were female. The calculation was made both for the number of researchers (this will be an over-estimate, as in Table 2 there are only 7 people, not 10) and for their total contributions.

The high percentage of females in China is clearly anomalous as fewer than half the names could be sexed – this was also the case for Taiwan and Korea. Among European countries, Canada and the USA, on average just over 80% of names could be sexed, and the female percentages are therefore more reliable. Austria, Belgium, Germany and the Netherlands score noticeably low on female participation. On the other hand Poland, a former Communist country where females were strongly encouraged to work (Webster, 2001), ranked highly, and the 10 other eastern European countries (the new "accession Member States" of the European Union) as a group ranked more highly still, with an actual majority of female researchers (51.5%) though their collective contribution was only 46.6%.

Table 3. Analysis of lung cancer researchers in different countries by sex. P = number of people; C = number of contributions (integer count). F = number of females; M = number of males. Countries are ranked by percentage of female researchers.

	Total			*Males*	*Females*	*Unknown*	*Sexed, %*		*F/(M+F), %*	
ISO	*P*	*C*	*C/P*	*P*	*P*	*P*	*P*	*C*	*P*	*C*
CN	13500	29897	2.21	2241	3918	7341	46	42	63.6	63.9
RoW	5226	8475	1.62	1920	1733	1573	70	74	47.4	45.8
PL	842	1643	1.95	396	348	98	88	91	46.8	43.2
IT	4647	9220	1.98	2060	1802	785	83	87	46.7	39.6
BR	721	911	1.26	338	282	101	86	86	45.5	43.9
ES	2300	4376	1.90	983	808	509	78	81	45.1	42.2
KR	3990	10533	2.64	938	754	2298	42	43	44.6	44.7
TR	1827	2747	1.50	819	648	360	80	83	44.2	39.0
SE	560	1159	2.07	268	205	93	84	86	43.3	39.7
TW	2867	8243	2.88	508	378	1981	31	34	42.7	38.5
Wld	36480	77204	2.12	10471	10876	15139	59	56	50.9	48.5
FR	3319	7976	2.40	1346	946	1027	69	80	41.3	38.2
DK	502	965	1.92	257	179	66	87	90	41.1	44.0
UK	2908	4782	1.64	1403	914	591	80	84	39.4	35.1
US	19962	44423	2.23	9854	6416	3692	82	84	39.4	34.9
AU	1101	2336	2.12	531	343	227	79	84	39.2	38.6
GR	1247	2194	1.76	620	369	258	79	85	37.3	31.1
CA	1933	4585	2.37	940	551	442	77	79	37.0	37.1
IN	940	1339	1.42	363	212	365	61	62	36.9	34.3
NO	300	923	3.08	172	95	33	89	93	35.6	26.2
NL	1638	3738	2.28	865	462	311	81	86	34.8	31.1
CH	756	1293	1.71	417	212	127	83	87	33.7	29.6
BE	606	1186	1.96	287	143	176	71	72	33.3	28.9
AT	412	851	2.07	242	105	65	84	89	30.3	23.1
DE	3523	6935	1.97	2083	841	599	83	88	28.8	23.9
JP	8900	24503	2.75	4260	1703	2937	67	68	28.6	22.1

The five South American countries (Argentina, Brazil, Chile, Colombia and Venezuela) also scored well for female participation with nearly 46% of researchers and 44% of contributions, slightly higher than the values for Brazil alone. The three Mediterranean Latin countries (Italy, Portugal and Spain) also scored well, and Portugal had the highest female participation, with over 61% of female researchers, whose contribution was 58%.

The correlation of the percentage of females in the above table (for the 11 countries for which a comparison could be made) with that obtained from another (unpublished) study on cancer screening where a similar methodology was used is quite high ($r^2 = 0.63$). However lung cancer averaged only 39% compared to 46% for cancer screening. Sweden was an exception, with a higher female percentage in lung cancer (43%) compared with 40% for cancer screening.

For the analysis of ethnicity/national origins of the researchers, we first determined the percentage of researchers with "own country" ethnicity. Table 4 shows, for each country, the national background(s) of the names that were selected and the corresponding percentages of their numbers and contributions.

Table 4. Numbers and percentages of "own country" researchers

Country	*Own CU*	*P, %*	*C, %*		*Country*	*Own CU*	*P, %*	*C, %*
BR	BR	26.4	27.1		NL	NL	62.9	63.8
DK	DK,SC	41.0	41.8		IN	IN	67.8	68.3
CA	FR,UK	42.0	42.9		ES	ES	68.3	67.3
SE	SC,SE	48.2	50.7		DE	DE	70.3	71.2
AU	UK	51.9	55.7		BE	BE,FR,NL	76.2	72.2
NO	NO,SC	55.3	58.8		TW	CN	78.9	74.5
FR	FR,UK	58.5	60.6		PL	PL	80.0	76.7
UK	UK	59.8	60.1		CN	CN	83.7	85.3
US	EUR	60.1	61.4		TR	TR	85.6	86.6
GR	GR	60.5	64.0		IT	IT	90.5	91.2
AT	DE	61.9	59.5		KR	KR	92.4	92.9
CH	DE,FR,IT	62.0	64.9		JP	JP	95.3	96.3

The result for Brazil is anomalous, as most of its researchers are descended from Europeans and would have European or Latin American names. (A scientific conference in Caxambu of the Brazilian Biochemical Society, which one of us attended in 1994, was almost entirely populated by Brazilians who appeared to be of European origin.) If these are allowed as "own country" names, then they would represent 90% of Brazilian researchers with a contribution of 91%.

The countries with the greatest fraction of their lung cancer workforce of non-native origin appeared to be the Nordic ones (Denmark, Sweden and Norway), and Canada. The UK also had a high proportion of its lung cancer researchers with non-national ethnic backgrounds (40%) and the same percentage of contributions. On the other hand, Italy had only 10% of non-Italians, and Korea and Japan even fewer foreigners (8% and 5% respectively) though there were rather more in Taiwan (21%) and in China (16%). This feature of Italian research was found in a previous study (Roe et al., 2014).

We now consider the contribution of other European researchers to the lung cancer research of the 14 selected European countries. This is shown in Table 5.

Table 5. Contributions of researchers from other European countries to the lung cancer research of 14 selected European countries. P = people; C = contributions (integer count).

	Other EUR, %				*Other EUR, %*				*Other EUR, %*	
Country	*P*	*C*		*Country*	*P*	*C*		*Country*	*P*	*C*
DK	52.4	53.8		FR	28.7	29.5		ES	17.3	19.9
NO	36.3	27.1		CH	27.4	25.4		BE	16.7	22.1
SE	35.7	36.1		NL	27.0	27.1		PL	16.4	19.5
GR	33.9	32.2		DE	21.5	21.2		IT	6.6	6.0
AT	33.7	37.4		UK	21.3	21.3				

The results are similar to those of Table 4, except that the UK dropped from fifth to tenth place with its proportion of other European nationals among its lung cancer researchers. Its acceptance of non-Europeans was therefore correspondingly greater. There were 7.0% with a South Asian background, three fifths of them Indian, 3.1% Chinese and 4.0% from other Asian countries. These percentages are much higher in Europe except that Sweden had a slightly greater percentage of researchers of Chinese origin. The UK also had 2.2% of lung cancer researchers with North African or Levantine names (third highest in Europe), 0.8% with African names (second to the Netherlands) and 0.7% with names from Latin America (highest in Europe). Altogether, its lung cancer research population with non-European names amounted to 19% of the total.

These percentages can be compared with census data for England and Wales in 2011 (ONS, 2012). There were about 5.3% of "other White" including Irish (corresponding approximately to "other Europeans" in the above table), 2.5% of Indian origin, 4.2% of other Asians, and 0.7% of Chinese. So the Chinese were over-represented among lung cancer researchers by 3.1/0.7 = 4.4, the Indians by 4.2/2.5 = 1.7 and other Asians were slightly under-represented by 4.0/4.2 = 0.95. The other Europeans were also over-represented by 21.3/5.3 = 4.0. Many of the Chinese would have been graduate students and would probably have returned to China or gone elsewhere after obtaining their doctorates or other degrees.

Canada and the USA were even more accepting of non-Europeans, and their percentages of the different groups are shown in Table 6. Almost 40% of US lung cancer researchers were of non-European ethnicity or national background, of whom by far the largest group were Chinese (13.8% of the total), followed by Indians (5.8%) and Koreans (3.5%). Despite the large numbers of Latin Americans now in the population, they represent only 4.3% of American lung cancer researchers, even when people with Brazilian, Portuguese and Spanish names are included. US Census data for 2010 show that "Latinos" accounted for well over one third of those living in the USA but born abroad, compared with the Chinese (5%) and Indians (4%). However, only 5% of them had university degrees, compared with 50% of the Chinese and 74% of the Indians (US Census Bureau, 2012).

Table 6. Percentages of non-European lung cancer researchers in Canada and the USA.

	CHI	*ASI*	*SAS*	*LEV*	*LAT*	*AFR*	*Other*	*Total*
CA	11.0	9.6	5.6	4.2	0.9	0.4	2.7	34.4
US	13.8	9.6	7.7	4.5	1.4	1.0	1.8	39.8

The file also allows us to determine where lung cancer researchers with given ethnicities are now based and how much they are contributing to either their countries of origin or their new host countries. We previously found (Basu, Roe & Lewison, 2012) that the output of cancer research papers by people of Indian origin now living in Canada and the USA was greater than that of Indians remaining in India. In lung cancer research, of the 2,233 researchers with Indian names, over half (1,164 or 52%) are working in the USA and only 637 (28.5%) in India. There are 124 in the UK, 80 in other European countries, 73 in Canada and 155 elsewhere. The situation is very different for the Chinese, Japanese and Koreans, see Table 7.

Table 7. Current locations of lung cancer researchers from China, Japan and Korea (S).

Ethnicity \ Workplace	*China*	*Europe*	*Japan*	*Korea*	*USA*	*Other*	*Total*
CN	11301	220	124	178	2762	2725	17310
JP	18	27	8485	9	341	90	8970
KR	1151	40	51	3688	702	443	6075
CN, %	65.3	1.3	0.7	1.0	16.0	15.7	
JP, %	0.2	0.3	94.6	0.1	3.8	1.0	
KR, %	18.9	0.7	0.8	60.7	11.6	7.3	

Clearly, most of these East Asians remain in their own country, although the Chinese travel abroad the most, and the Japanese the least, and hardly at all to China or Korea. There is also very little movement to Japan by Chinese and Koreans, and some of the 51 Koreans working in Japan may be ones whose families have been there for several generations. In 2005, there were some 901,000 people of Korean ancestry living in Japan (out of a population of 128 million) or 0.7%. The percentage of the lung cancer researchers in Japan with Korean names was 0.6%, which is slightly less.

We can also see where the lung cancer researchers with various "European" names are now - some will have stayed in their own country, some have gone to the United States, and some have gone elsewhere. The two figures below show the situation. The five largest countries (in terms of numbers of named researchers) are on the left chart and the next nine are on the right chart. However, many of those with British, German, Polish and Irish names will have been resident in the USA for several generations rather than being recent immigrants.

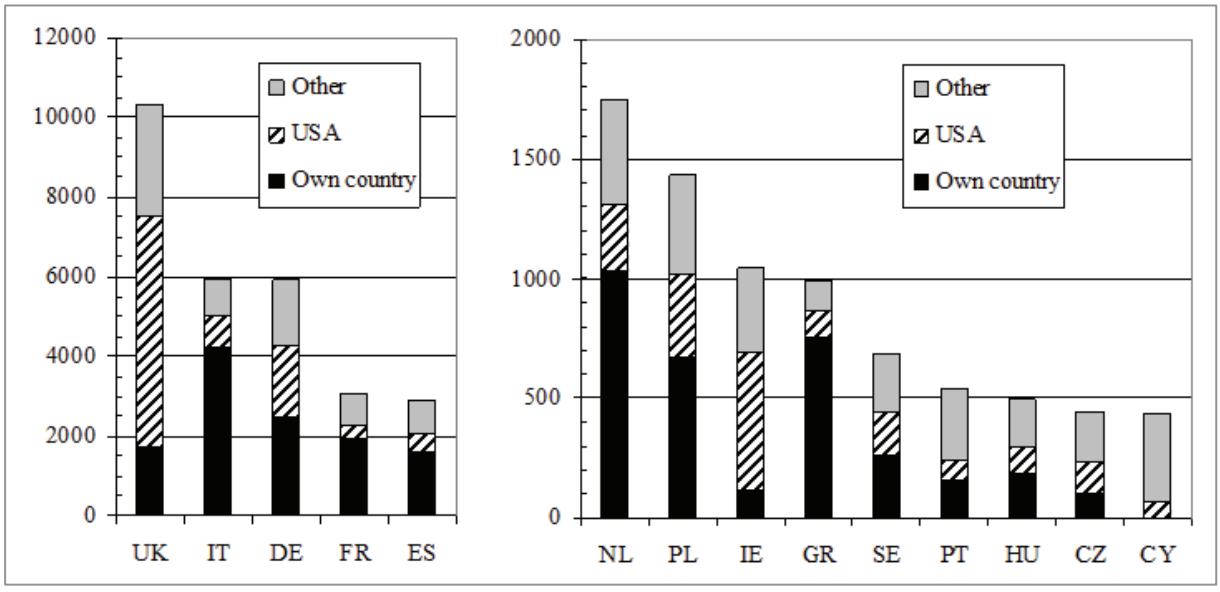


Figure 1. Locations of lung cancer researchers with names characteristic of different European countries - in own country, in the USA, and in other countries.

The file of lung cancer researchers also enables us to investigate whether there is a difference between men and women in the numbers of papers that they write. Figure 2 shows the sex ratio F/(M+F) for groups of authors who publish sufficient papers to put them in a given centile. Thus of the 84,533 authors, the top 1% (n = 845) each wrote at least 17 papers, and the figure shows that just under 26% of those whose sex could be determined were female. By contrast, the 53,143 authors with but a single paper (probably mainly graduate students) were nearly 44% female. This shows clearly that the percentage of females falls off with production, which is probably strongly correlated with seniority. A similar graph could be

produced for individual countries, or ethnic groups, provided that there are enough people in the group or country to make the analysis worth-while.

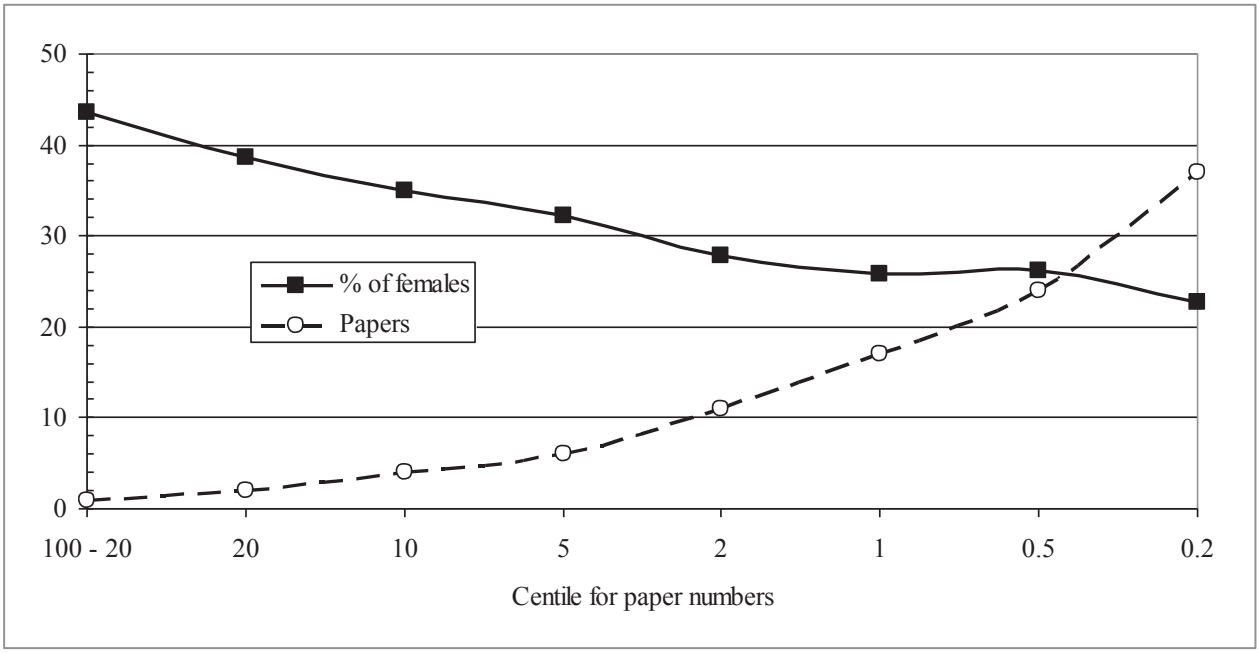


Figure 2. Percentage of female authors whose number of lung cancer papers put them in given centiles of the population of 84,533 authors.

Discussion

This paper greatly extends the methodology used in Roe *et al*., 2014 by its application to all the papers in a subject area, including multi-national ones, and by the provision of a file of all the named researchers, classified by their ethnicity and sex, and the country or countries in which they were working. This allows many research questions to be addressed, and some of them have been in this paper.

However, the methodology still has some limitations, and these are currently being tackled. The first is that, although Aakre, J. can be identified as the same as Aakre, Jeremiah and so classed as male, the file contains two separate entries (actually three in this case because he also published a paper with a Chinese address), which should be amalgamated. The second limitation is that the number of each researcher's papers is given only as an integer count, and for many purposes it would be more useful to have a fractional count, based on the number of different authors of each paper. This is sometimes problematic, as quite a lot of papers list individuals with more than one affiliation. This would not matter if these are all in the same country, as is usual, but increasingly nowadays senior researchers have appointments in more than one country. We would need to fractionate these people's contributions by country in order to make the sum of the individual contributions equal the number of papers (less those with anonymous authors).

A further problem is that, although most names can be classed by country or region within it, some can not be, at present. (The lung cancer database only has 392 names not classified by ethnicity, less than 0.5% of the total.) This is well within the margin of error for most bibliometric studies. However, there is a bigger problem with ambiguous family names where the given names are not on the paper. We have approached this on the basis that most East Asians stay in their own country (see Table 7). However this method would not apply so strongly to Europeans, and as movement and marriage between EU Member States becomes increasingly common, there will be more errors in attribution of researchers to countries.

We have also found that the percentage of names that cannot be sexed is quite high, so that the results for some countries are not at all representative – notably for China. Clearly, we need to acquire more information on the sex associated with particular Chinese, Japanese and Korean names, although some names may not be strictly unisexual. (This occurs also with some European and some British given names, such as Hilary and Robin, where a minority of holders are respectively male and female.) We previously took a ratio of at least 10:1 as indicative of the association of a given name with just one sex, but there may be some errors, though these could be reduced if a researcher has two given names and one can be sexed definitively. This again will need improvements to the software.

Acknowledgments

This study was funded by King's Health Partners and the Global Lung Cancer Coalition as part of their evaluation of lung cancer research world-wide.

References

Basu, A., Roe, P. & Lewison, G. (2012) The Indian diaspora in cancer research: a bibliometric assessment for Canada and the USA. *Proceedings of the 17th International Conference on Science and Technology Indicators* (eds. Éric Archambault, Yves Gingras and Vincent Larivière), Montréal: Science-Metrix and OST; 110-120.

Bettachy, A., Baitoul, M., Benelmostafa, M. & Mimouni, Z. (2013) Women in scientific research in physics in Morocco. *Women in Physics*, *1517*, 128-129.

Campbell, A.G., Leibowitz, M.J. Murray, S.A., Burgess, D., Denetclaw, W.F., Carrero-Martinez, F.A. & Asai, D.J. (2013) Partnered research experiences for junior faculty at minority-serving institutions enhance professional success *CBE-Life Sciences Education*, *12*, 394-402.

Garg, K.C. & Kumar, S. (2014) Scientometric profile of Indian scientific output in life sciences with a focus on the contributions of women scientists. *Scientometrics*, *98*, 1771-1783.

Gonenc, I.M., Akgun, S., Bahar Ozvaris, S. & Emin Tunc, T. (2013). An analysis of the relationship between academic career and sex at Hacettepe University. *Egitim ve Bilim-Education and Science*, *38*, 166-178.

Griffin, K.A., Bennett, J.C. & Harris, J. (2013). Marginalizing merit? Gender differences in black faculty discourses on tenure, advancement, and professional success. *Review of Higher Education*, *36*, 489-512.

Hassouneh, D., Lutz, K.F., Beckett, A.K., Junkins, E.P. & Horton, L.L. (2014). The experiences of underrepresented minority faculty in schools of medicine. *Medical Education Online*, *19*:24768. doi: 10.3402/meo.v19.24768.

Homma, M.K., Motohashi, R. & Ohtsubo, H. (2013). Maximizing the potential of scientists in Japan: promoting equal participation for women scientists through leadership development. *Genes to Cells*, *18*, 529-532.

Isfandyari-Moghaddam, A. & Hasanzadeh, M. (2013). A study of factors inhibiting research productivity of Iranian women in ISI. *Scientometrics*, *95*, 797-815.

ONS. (2012). http://www.ons.gov.uk/ons/dcp171776_290558.pdf

Pololi, L.H., Evans, A.T., Gibbs, B.K., Krupat, E., Brennan, R.T. & Civian, J.T. (2013). The experience of minority faculty who are underrepresented in medicine, at 26 representative US medical schools. *Academic Medicine*, *88*, 1308-1314.

Roe, P., Lewison, G. & Webber, R. (2014). The sex and ethnicity or national origins of researchers in astronomy and oncology in four countries, 2006-2007 and 2011-2012. *Scientometrics*, *100*, 287-296.

US Census Bureau. (2012). https://www.census.gov/newsroom/pdf/cspan_fb_slides.pdf.

Webster, B.M. (2001). Polish women in science: a bibliometric analysis of Polish science and its publications, 1980-1999. *Research Evaluation*, *10* (3), 185-194.

A Model for Publication and Citation Statistics of Individual Authors

Wolfgang Glänzel[1, 2], Sarah Heeffer[1], and Bart Thijs[1]

[1] *{wolfgang.glanzel, sarah.heeffer, bart.thijs}@kuleuven.be*
[1]KU Leuven, ECOOM and Dept. MSI, Leuven (Belgium)
[2] Library of the Hungarian Academy of Sciences, Dept. Science Policy & Scientometrics, Budapest (Hungary)

Abstract
One of the most important requirements of building applicable models and meaningful indicators for the use of scientometrics at the micro and meso level is the correct identification and disambiguation of authors and institutes. Platforms like ResearcherID or ORCID with author registration providing high reliability but lower coverage now provide appropriate data sets for the development and testing of stochastic models describing the publication activity and citation impact of individual authors. This paper proposes a triangular model incorporating papers, citations and authors analogously to the dichotomous model used at higher levels of aggregation like countries or fields. This model is applied to a set of authors in any field of science identified by their ResearcherID. However, the main advantage of classical citation indicators to study citation impact under conditional productivity turned out to be the main problem in this triangle: the possible heterogeneity of the collaborating authors results in low robustness. A mere technical solution to this problem would be fractional counting at three levels, but the conceptual issue, the different roles of co-authors causing this heterogeneity, will never be solved by any algorithm.

Conference Topics
Methods and techniques; Data Accuracy and disambiguation

Introduction
Spectacular progress has been made in author identification, the disambiguation of names and their institutional assignment on the basis of correct affiliation and cleaned address data extracted from bibliographic databases. In particular, this is one of the most important and basic requirements of building applicable models and meaningful indicators for the use of scientometrics at the micro and meso level. Correct author identification is not only indispensable in studies of academic careers, researchers' mobility, authors' publication and collaboration patterns (Braun et al., 2001) but also in monitoring constitution and performance of research teams (Strotman & Zhao, 2012). The task outlined here is practically twofold: On the one hand, the large-scale disambiguation and assignment of authors forms still one of the big challenges in scientometrics. Although the quality of disambiguation and assignments of authors has considerably improved due to sophisticated algorithms and scientometric techniques, e.g., using "bibliometric fingerprints" (Tang & Walsh, 2010) and similarity patterns (cf. Caron & van Eck, 2014), automated processes proved not sufficient to provide reliable reference standards even if optional interaction of individual authors has been made possible. In this context author identification of the Mathematical Reviews and Elsevier's Scopus databases might just serve as examples. Mathematical Reviews was one of the first databases that applied automated processes (since 1985) for author identification. Challenges are, among others, mobility, topic shifts, career breaks, occasional and infrequent publication activity, e.g., so-called transients (Price & Gürsey, 1976). Incorrect institutional assignment, multiple identities as well as unresolved homonyms are still frequently observed errors. This is contrasted by the possibly higher reliability but lower coverage of identifiers that are based on author registration as, for instance, the ResearcherID of the Web of Science database (Thomson Reuters) and the Open Researcher and Contributor ID (ORCID). The latter IDs are sensitive to human errors and their willingness to regularly update and maintain publication assignment to their IDs. A previous study has pointed to the representativeness bias in favour of more prolific authors (Heeffer et al., 2013).

The second issue is partially related to methodology but also of conceptual nature. The methodological issues arise from the superposition of multiple assignments of publication to subjects, on the one hand, and to co-authors and their particular profiles, on the other hand.
Stochastic models for publication activity and citation impact of authors, however, require partitions, which can only partially approximated by corresponding fractionation procedures (cf. Glänzel et al., 2014 for multiple subject assignment in the context of Characteristic Scores at Scales at different levels of aggregation). A further issue arises from the different stages of the individual careers of authors at the same time; while the same publication year ensures the same age of papers in a given citation window, a pre-set publication year collects papers of scientists who are situated in completely different stages of their careers at the same time. The fact that a PhD student or post-doctoral fellow might collaborate with a senior scientist makes the situation even more complex. Thus the question arises whether the same reference standard derived from the data set should apply to the junior as to the senior co-author. And this leads us directly to the conceptual problem: What is the weight of co-authors and their profiles in determining standards for possible benchmarking exercises? This implies that large-scale statistics calculated on the basis of given publication periods and selected subject fields will not be appropriate as reference standards at the micro level but might indeed mirror the profiles of larger institutions and countries adequately and thus serve as general model at these levels of aggregation.
In this paper a triangular stochastic model analogously to the models used at higher levels of aggregation will be described and opportunities and limitations of such a model will be discussed. In the following we will mainly focus on the following questions.

1. What is the relationship between authors' productivity and their citation impact?
2. How can the relationship between the authors' citation impact and the impact of their publications be described?
3. What is the possible effect of co-authorship on these patterns?
4. Can any reference standard for evaluative studies be derived from the model and the empirical data?

This short introduction already adumbrates the possibilities but also the limitations of scientometric models that are created on the basis of the identification and assignment of individual authors. We optionally attempted to use Thomson Reuters' *Distinct Author Identification System* (DAIS), which is based on clustering author names, institution names, and citing and cited author relationships (Thomson Reuters, 2012). As all automated processes, this results in a broader coverage, but suffers from false positives. We have found nearly 30 authors with more than 300 papers each in 2011 according to the DAIS and the most productive author had 1272 WoS indexed papers. However, a simple manual check of names and profiles of authors associated with the same DAIS code revealed different persons with the same family name and first initial but partially different given names and different research profiles. In order to reduce uncertainty we decided therefore to use Thomson Reuters' ResearcherID in conjunction with journal articles published in the same year hazarding the consequences of representativeness bias. From the viewpoint of the model and the analysis this restriction is, however, immaterial. In this context we would like to stress again that the possible biases in representativeness of author selection is insignificant from the viewpoint of the creation and applicability of the model. More important in this context is the reliability of identification of the authors and their affiliations. Nevertheless, we will first have a look at representativeness of author selection on the basis of Thomson Reuters' ResearcherID (RID). This first part of the analysis forms a straight continuation of a previous study on productivity of registered authors by Heeffer et al. (2013).

Data sources and data processing

All papers indexed as articles, proceedings papers, reviews and letters in the 2011 volume of Thomson Reuters Science Citation Index Expanded (SCIE) have been selected. The reason for this choice of a single year publication window, which results from structural properties of author representation and productivity reflected by annual document indexation in bibliographic databases, is as follows. We have already mentioned at the outset that citation processes of scientific papers published in the same year have the following properties: Within a given citation window, all documents in the set have the same age at any particular time and the citation process is not *homogeneous*, that is, citation frequencies at the initial period differ from those at later stages. Paradigmatically this phenomenon has been characterised as a combination of phases of maturing and decline in citation processes (Glänzel & Schoepflin, 1995; Moed et al., 1998). As a consequence, enlarging the citation window will not simply result in a multiplication of citations by a factor proportional to the length of the window. The situation is completely different when a population with heterogeneous age structure is underlying the process and authors are constantly entering and leaving the system. While the citation process of a fixed document set can be described, for instance, by a simple birth process (e.g., Glänzel & Schoepflin, 1994), the publication distribution of an author set, which is subject to changes and interacts with the "environment", requires a different model taking also the effect of immigration and emigration into account. Such model has been proposed by Schubert and Glänzel (1984). This is the situation we find in any publication period in a bibliographic database: Newcomers are entering the author population, terminators are leaving the system and continuants are members of the population for a longer time including the complete period under study (cf. Price & Gürsey, 1976). As a consequence, publication activity in a longer time period can be simulated by multiplying productivity by a proportionality factor according to the length of the period. Therefore it is initially sufficient to select a shorter period of, e.g., one year as the basis of the analysis.

The reason why we have chosen the year 2011 was that in this particular year the share of papers with registered RID was the largest. We expected, of course, that this share will increase and that more authors will be registered in more recent years but the fact that this share decreases beyond 2011 is probably caused by the attitude of authors to update registration and register newly indexed papers not always immediately and regularly but rather intermittently. The choice of 2011 was also convenient because it allows the observation of citations in an appropriate time span. In addition to this publication year we could therefore choose the three-year citation window 2011-2013.

Methods and results

Theoretical considerations

As already mentioned in the previous section, the inclusion of productivity patterns in citation statistics permits insight into a complex system with the provision of a whole set of benchmarks and reference values. From the mathematical viewpoint, we deal with two basic variables that can stochastically be considered random variables, ζ expressing publication activity and ξ standing for citation rates. Yet the two variables are not assumed to be independent and it is commonly known that more prolific authors tend to be more cited as well. Therefore $P(\xi=i|\zeta=j)$ does not necessarily equal $P(\xi=i)$ for all $i, j \geq 0$ and the conditional expectation $E(\xi|\zeta=j)$, being a function of ζ and taking its values with probability $P(\zeta=j)$ is not necessarily constant. In our case, the following measurable variables occur: The publication activity of a (randomly chosen) author in the mirror of the SCIE database in 2011, the citation impact of a (randomly chosen) paper indexed in the 2011 volume of the SCIE and the citation impact of an author with one or more papers in 2011 with the intermediate conditional

measure of citation impact, provided the author has a given number of publications $j \geq 0$ in 2011.

The following mathematical description, which is indeed necessary to avoid confusions, will, however, be restricted to the absolute necessary. The first question formulated in the introduction relates to the relationship between authors' productivity and their citation impact. This can be formulated as follows. Since citation impact is always measured through the citation rates of individual publications, an author's citation impact can theoretically be obtained as

$$P(\xi=i) = \Sigma_j P(\xi=i|\zeta=j)\cdot P(\zeta=j) \text{ for all } i \geq 0,$$

with the corresponding expectation

$$E(\xi) = \Sigma_j E(\xi|\zeta=j)\cdot P(\zeta=j).$$

Index j is assumed to be positive because the trivial case $P(\xi=i|\zeta=0) = 1$, if $i = 0$ and $P(\xi=i|\zeta=0) = 0$, otherwise, can be excluded (no citations without publications). The corresponding statistics are then denoted as $f_i|j$ and $x|j$. Both statistics (conditional empirical distribution and mean value) refer to the citation impact of authors. Furthermore, the corresponding conditional mean citation rate of an author's *papers* can be obtained by dividing $x|j$ by the number of papers j, that is, $(x|j)/j$ with $j > 0$ is an estimator of the expected citation rate of the individual papers of an author with j papers in the given publication year.

In order to tackle the second problem, we have to introduce a third variable, which will complete the triangular model. Using the notation η for the citation impact of a single paper by an individual author, we obtain a more complex formula than above for the conditional probabilities taking all possible combinatorial combinations concerning number of publications and their citations into account but the relationship of their expectations simply reduced to $E(\xi) = E(\eta)\cdot E(\zeta)$. Under the simple assumption that the likelihood not to be cited is the same for all papers of the author, i.e., $q = P(\eta=0)$ for all $j > 0$, we can approximate the probability of author uncitedness and citedness as $P(\xi=0) = \Sigma_j q^j \cdot P(\zeta=j) = P(\eta=0)^j$ and $P(\xi>0) = 1–P(\xi=0)$, respectively. The reason for the relative simplicity of this expression is that uncitedness of an author in a given period implies that none of his/her papers is cited. The extreme cases $P(\xi=0) = 0$ and $P(\xi=0) = 1$ are obviously equivalent with $q = 0$ and $q = 1$, respectively. We will denote the empirical value of q by g_0. Using the mean values x, z and y as estimators of expected citation rate of an author, the expected publication activity of an author and the expected impact of the author's papers, respectively, we obtain the simple relationship $x = y\cdot z$. From the elementary considerations we can conclude that at least basic statistics can be readily expressed with the aid of two variables.

Finally, it might be worth mentioning in this context that the above random variables and the corresponding statistics also form the groundwork for modelling Hirsch-type indices, notably their cumulative versions such as the successive h-index (e.g., Schubert, 2007).

The sample

The sample of RID authors does – as already observed by Heeffer et al. (2013) – not form a *random sample* of the complete author population in the database as RID authors are less frequent at the low end (particularly among single-paper authors), and are more productive at the high end of the productivity distribution.

Table 1. Share of papers with RID authors and their relative citation impact by countries [Data sourced from Thomson Reuters Web of Science Core Collection].

Country	*Papers*	*RCR*	*NMCR*	*%HC*	*RCR*	*NMCR*	*%HC*	*%RID*
Argentina	7702	1.03	0.98	1.3%	1.44	1.91	4.5%	14.7%
Australia	40979	1.16	1.36	2.1%	1.22	1.63	2.9%	42.4%
Austria	12274	1.22	1.45	2.6%	1.39	2.01	4.7%	29.7%
Belgium	17598	1.22	1.51	2.5%	1.34	1.96	4.1%	32.6%
Brazil	33940	0.99	0.72	0.7%	1.02	0.88	1.0%	45.2%
Canada	54511	1.14	1.38	2.1%	1.38	2.08	4.3%	21.0%
Chile	5073	1.15	1.08	1.3%	1.31	1.49	2.6%	31.8%
Czech Rep.	9350	1.18	1.09	1.5%	1.27	1.40	2.4%	40.4%
Denmark	12772	1.30	1.62	3.1%	1.41	2.03	4.4%	36.2%
Egypt	6251	1.02	0.75	0.6%	1.41	1.52	2.9%	15.1%
Finland	9945	1.20	1.42	2.2%	1.35	1.91	3.8%	34.7%
France	65238	1.09	1.29	1.8%	1.20	1.71	3.0%	28.4%
Germany	91263	1.14	1.39	2.1%	1.23	1.81	3.4%	30.7%
Greece	10647	1.13	1.12	1.6%	1.45	1.91	4.1%	22.2%
Hungary	5763	1.15	1.16	1.8%	1.36	1.63	3.4%	36.2%
India	46532	0.98	0.68	0.7%	1.20	1.26	1.8%	13.0%
Iran	20234	1.15	0.71	0.8%	1.55	1.36	2.8%	9.1%
Ireland	6833	1.18	1.42	2.3%	1.34	1.85	3.5%	35.5%
Israel	11558	1.06	1.34	2.1%	1.28	1.97	4.2%	21.4%
Italy	53919	1.10	1.22	1.7%	1.19	1.52	2.6%	32.8%
Japan	76799	0.94	0.96	1.1%	1.13	1.52	2.5%	20.9%
Malaysia	7325	1.12	0.71	0.7%	1.15	0.84	0.9%	41.1%
Mexico	9830	1.02	0.89	1.2%	1.40	1.69	3.4%	21.0%
Netherlands	31883	1.21	1.60	2.8%	1.28	1.90	3.8%	36.8%
New Zealand	7186	1.17	1.33	2.1%	1.45	1.98	4.0%	30.5%
Norway	9694	1.23	1.43	2.4%	1.43	2.07	4.8%	26.7%
Pakistan	5371	1.18	0.69	1.1%	1.52	1.58	3.3%	16.0%
China PR	156403	1.04	0.91	1.1%	1.24	1.53	2.9%	20.2%
Poland	20261	1.08	0.82	0.9%	1.30	1.41	2.3%	20.1%
Portugal	9844	1.14	1.19	1.6%	1.17	1.29	1.9%	63.9%
Romania	6618	1.26	0.71	1.2%	1.30	0.97	1.9%	40.0%
Russia	27853	1.03	0.55	0.7%	1.12	0.94	1.5%	26.5%
Saudi Arabia	5417	1.15	0.92	1.3%	1.35	1.42	2.4%	31.4%
Singapore	9458	1.17	1.53	2.8%	1.29	1.91	4.1%	47.0%
South Africa	7787	1.26	1.19	2.2%	1.50	1.73	4.4%	25.3%
South Korea	44228	0.97	0.89	1.0%	1.13	1.44	2.4%	22.2%
Spain	47885	1.10	1.24	1.7%	1.19	1.56	2.6%	35.9%
Sweden	19923	1.18	1.44	2.4%	1.31	1.90	3.8%	30.8%
Switzerland	23582	1.29	1.73	3.3%	1.38	2.16	5.0%	34.6%
Taiwan	25550	0.92	0.93	1.1%	1.19	1.55	2.9%	17.0%
Thailand	5819	1.08	0.89	1.0%	1.32	1.48	2.6%	16.9%
Turkey	22571	1.02	0.63	0.8%	1.39	1.34	2.7%	12.8%
UK	91438	1.16	1.46	2.4%	1.28	1.90	3.9%	31.6%
USA	333610	1.09	1.40	2.2%	1.25	1.95	3.9%	20.0%
World total	1229248	1.00	1.00	1.2%	1.13	1.42	2.2%	21.1%

Nevertheless, from the viewpoint of the objectives of this study, this bias is primarily insignificant. In total we have 1,229,248 documents among which 259,341, that is, 21.1% had

at least one registered (RID) author. This share considerably varies among countries. The share ranges between about 10% in Africa, Arabic countries and India till about 50% and even more in Brazil, Singapore and Portugal.

Table 1 displays statistics of countries with at least 5,000 publications in 2011. In particular, the variable RCR represents the relation of observed citation impact and the corresponding journal-based expectation, NMCR stands for corresponding relation between observation and discipline-based expectation and %HC is the share of highly cited papers, that is, of papers that have received at least seven times as many citations as the standard of their discipline (see Glänzel et al., 2009 for exact definitions). The last variable %RID, finally, expresses the share of papers with (at least one) author with registered RID. The comparison of relative citation rates and the share of highly cited papers provides empirical evidence that papers by registered authors exhibit distinctly higher citation impact than the corresponding national standards. We would also like to mention that only very few exceptions have been found in smaller countries not displayed here, e.g., Jordan and Latvia, where the share of highly cited papers and the RCR values did not reach their national standards created by all authors.

Representativeness of publications by authors with RID in individual subject fields is in line with our intuitive expectations: The share of papers by RID authors is the lowest in Mathematics (13.0%), clinical and experimental medicine (14.2% for general & internal medicine and 14.2% for non-internal specialties) and engineering (18.7%). This is contrasted by the corresponding shares in physics, chemistry and biosciences (29.8%, 28.5% and 25.0%, respectively).

Productivity and impact of RID authors

The bias in publication-activity statistics of registered authors has already been stressed (cf. Figure 3 in Heeffer et al, 2013). In particular, RID authors are less frequent at the low end, and more productive at the high end of the productivity scale. Figure 1 shows the distribution of papers over RID authors in 2011. The underlying data are based on the short period of only one year so that the share of single-paper authors is consequently large. Nevertheless, the productivity distribution has the expected long tail: 87 authors have (co-)authored more than 50 papers each. We just mention in passing that the maximum count amounted to 296. This almost incredibly large annual publication output of publishing almost one paper a day is, however, formally correct. The author with an affiliation at the University Sains in Malaysia and a second, more recent one at the King Saud University in Saudi Arabia is active in crystallography. In this context we have to notice that the number of his co-authors per paper is rather low, so that even fractionation would not essentially decrease this author's publication count. This example also illustrates that conceptual issues might have more weight than the number or seniority of co-authors. Before we discuss field-specific aspects of authorship statistics, we still have a look at general citation patterns.

In Figure 2, the citation distribution over authors is compared with the corresponding distribution by papers. In addition to the two series of bars expressing the frequency of citations by RID authors and their papers, respectively, a solid line displays the citation distribution of all papers indexed in the SCIE database to illustrate the bias of the sample. The more moderate skewness and greater expectation of the distribution of citations over authors are plausible and in line with the theoretical rudiments described in the previous subsection since usually we have $z > 1$ and $g_0 \in (0, 1)$.

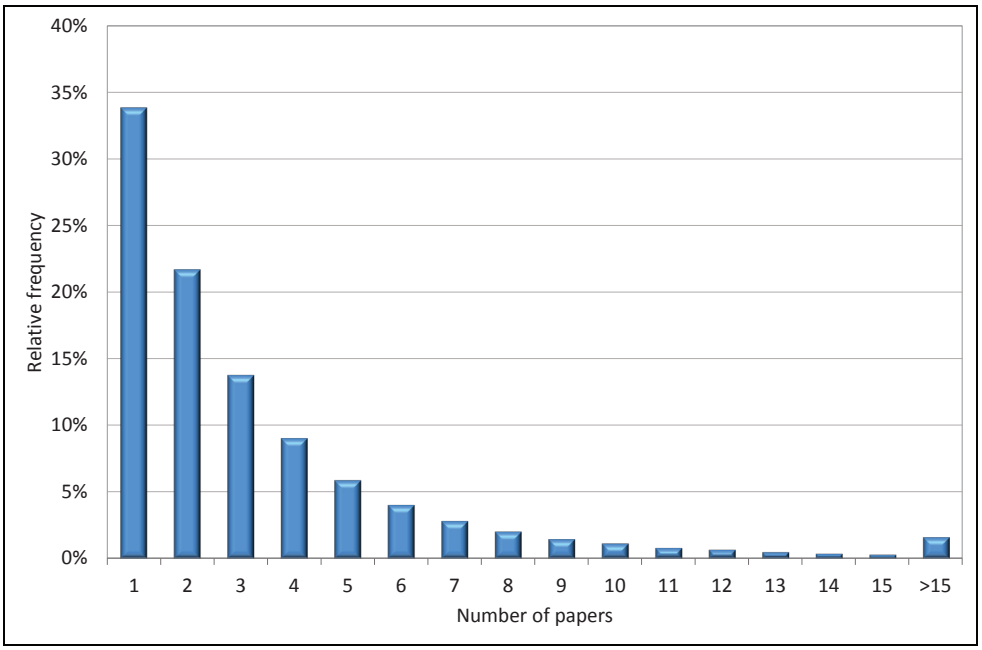


Figure 1. Relative frequency of publication activity of RID authors in 2011. [Data sourced from Thomson Reuters Web of Science Core Collection].

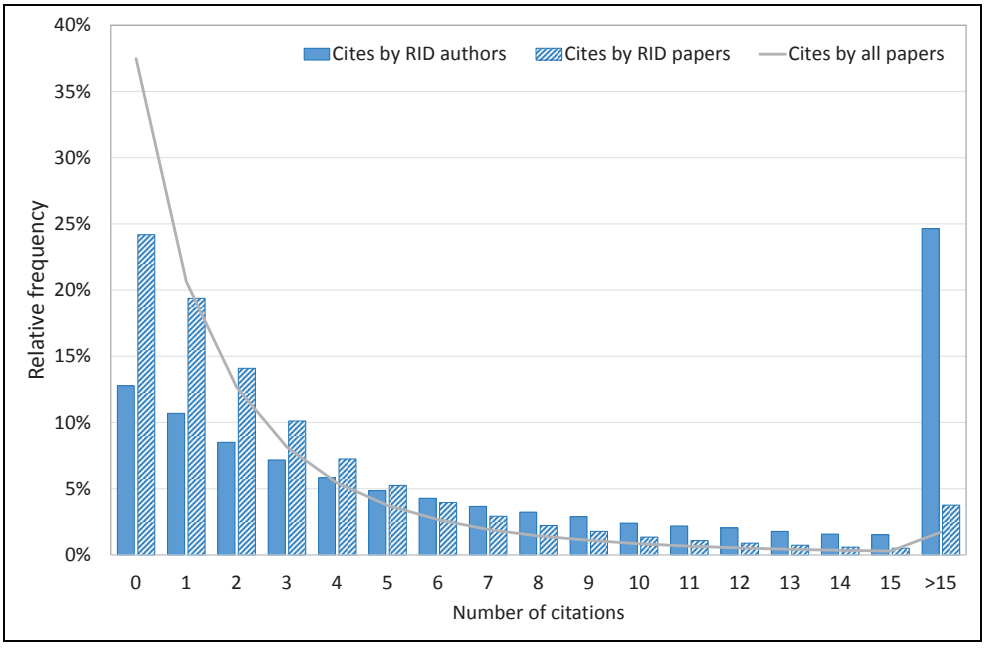


Figure 2. Empirical citation distribution related to RID authors in 2011 in a 3-year citation window. [Data sourced from Thomson Reuters Web of Science Core Collection].

A simple regression analysis aims at studying the relationship of productivity and citation impact of authors, on the one hand, and his/her publications, on the other hand. Conditional mean citation rates in the citation window 2011–2013 received by papers published in 2011 by registered authors have been plotted against their productivity (see Figure 3). Productivity higher than 32 papers has been omitted because of low frequency and considerably fluctuations beyond this level. A power-law model for author citations reflects a very strong correlation, whereas the regression for article citations by authors proved to be linear with somewhat weaker correlation.

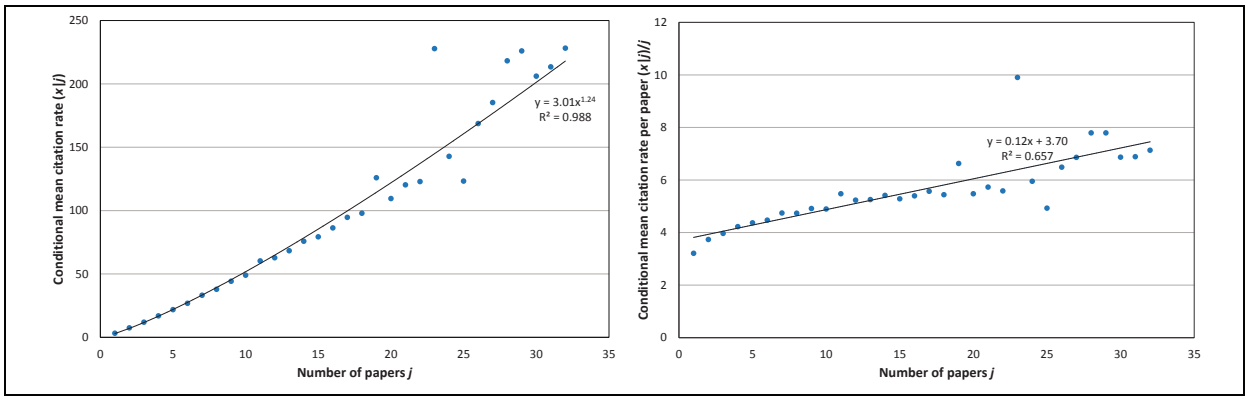


Figure 3. Plot of conditional citation impact of RID authors (left-hand side) and RID papers (right-hand side) based on a 3-year citation window vs. productivity in 2011 [Data sourced from Thomson Reuters Web of Science Core Collection].

While a positive effect of productivity on the expected citation impact of authors was, of course, expected (an increase of papers cannot result in less citations), the positive correlation between number of papers and the mean citation rate of *those* papers is as such not necessarily an inherent property of the model and as we described in the first subsection, the three variables ξ, ζ and η are not assumed to be independent. This indeed substantiates that the publication output of more productive authors exhibit also higher mean citation rates of their output. We have to emphasise that this holds at least for registered authors.

Table 2. Indicators of productivity and citation impact of RID authors and their papers by major science fields. [Data sourced from Thomson Reuters Web of Science Core Collection].

Field	*y*	*z*	*x*	f_0	g_0
A	2.76	1.32	3.65	17.9%	26.4%
Z	3.26	1.40	4.57	14.6%	22.8%
B	4.85	1.19	5.78	11.6%	15.9%
R	3.56	1.15	4.10	16.7%	22.0%
I	4.83	1.58	7.62	13.0%	20.9%
M	3.01	1.76	5.29	16.8%	27.8%
N	3.75	1.54	5.78	14.0%	20.8%
C	4.27	1.89	8.08	12.9%	20.8%
P	3.56	1.66	5.91	14.7%	25.1%
G	3.85	1.40	5.39	15.1%	20.9%
E	2.19	1.35	2.96	26.0%	36.4%
H	1.52	1.47	2.23	35.5%	44.6%

* 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; P: physics; R: biomedical research; Z: biology (organismic & supraorganismic level)

In order to conclude the analysis, we have calculated the mean values of the basic statistics *x*, *y* and *z* as well as the shares of cited authors and papers f_0 and g_0 by subject fields (see previous subsection for description). Table 2 shows these indicators for the 12 major fields in the sciences according to the Leuven–Budapest classification scheme (see Glänzel & Schubert, 2003). As explained in the theoretical part $x = y \cdot z$, $x \geq y$ and $f_0 \leq g_0$ is to be observed. Also subject-specific peculiarities are expected. The *y* and g_0 values concerning the citation impact of papers are by and large in line with the expectations: high impact and low

share of uncited papers in the biomedical sciences and the opposite situation in engineering and mathematics. Nevertheless, the very high impact of chemistry (with low uncitedness) was somewhat surprising and somewhat deviates from the general citation patterns of the fields. Chemistry seems also to be somewhat overrepresented in terms of author registration; 33.5% of all RID authors are active in this field. This is followed by physics with 27.4% and biosciences with 20.8%. All other fields have shares of registered authors below 20% with neuroscience and mathematics having the lowest ones (7.6 % and 4.4%, respectively). In this context we have to mention that the distribution of shares of RID authors over fields is rather strongly correlated with the corresponding distributions of their papers ($r = 0.928$). Hence the question arises whether statistics as presented in Table 2 could be used as reference standards for publication activity and citation impact of authors at the national or institutional level. It has already be stressed in the introduction that an application at the individual level is not recommended because of the heterogeneous age and profile structure of the underlying reference data. Other details regarding this question will be tackled in the following subsection.

Limitations

After the methodological groundwork has been laid for capturing and describing the relationship between productivity and citation impact of authors and their papers, we have also to look at considerable limitations of possible applications of the indicators derived from this model. The low variation of average productivity over subject fields gives already a first hint of possible issues. As already observed by Heeffer et al. (2013) on the basis of the three-year publication period 2009–2011 and RID authors from eight selected countries, the distribution of average productivity was rather flat and ranged – except for physics – roughly between 2 and 3 papers by RID author. Only the average activity in physics with 5 papers per author was distinctly higher. The accustomed and specific inequality of citation impact of papers in different subject areas is almost missing in the productivity statistics what surprises since it is known that scientists in mathematics and engineering are usually less productive – at least as reflected by journal literature – than their colleagues in most fields of the natural and above all in the life sciences. The reason for the observed phenomenon is quite complex but readily explicable. In order to discuss this in detail we have first to refer to the corresponding statistics on citation rates of given paper sets. Provided that the publication year or period as well as the citation window is properly defined and chosen and the subject classification is appropriate, multiple subject assignment of individual papers is then the only severe issue to cope with. Various fractional counting and weighting models have been developed to overcome this problem and to build suitable reference standards for benchmark analysis. Even for more complex statistics than simple shares and means, fractionation by subject can still yield extremely robust statistics as the methods of characteristic scores and scales has shown for various citation windows and aggregation levels (cf. Glänzel, 2007; Glänzel et al., 2014). The question of co-authorship, in general, and how the individual co-authors' actual contribution to a paper should be credited, in particular, is at least in the context of paper-based citation indicators a secondary issue and not primarily related to the definition of citation indicators. The situation becomes completely different, whenever author productivity is directly included in indicator building as, for instance, in our "triangle model" based on the author-paper-citation relationship. The different (academic) age and the different profiles of authors have already been mentioned as possible sources of bias or even distortion, notably in the context of creating benchmarks for individual-author statistics. The most serious issues are related to co-authorship and cannot be simply solved by fractionation by co-authors and/or subjects. Collaboration of senior with junior co-authors, that is, of authors with strong publication record and less active authors, independently of their actual contribution to

the paper in question and their function in preparing it, might have quite strong effect on the resulting indicators at the author level but also at higher level of aggregations. Here we would also like to point to two further issues, firstly the fact that a prolific author in one subject might only play a marginal part as researcher in a different subject in which he/she is collaborating with a possibly less prolific author, who, however, takes the part of the senior co-author of the paper(s) in this topic. Secondly, when it comes to measuring citation impact, an uncited author might be a co-author of a frequently cited author but the joint publications are not cited. This also implies that a mere author–citation analysis in conjunction with productivity studies does not yet suffice; an additional paper–citation analysis is needed for an adequate interpretation. And it becomes clear that a simple fractionation algorithm will not be able to solve these problems. A superposition of fractional counting at three levels (co-author credit, assignment by author profile and subject of publications) is required to solve at least the technical part of this problem: the large overlap by multiple assignments (authors, papers, subjects) could, of course, be resolved and indicators could then be additive over these actors and units at the price of very low robustness. Finally, the most important conceptual issue described in this subjection, the different roles of authors in different environments, will never be solved by using any algorithm.

Concluding discussion

Elementary statistics including relative frequencies and (conditional) mean values have been used to illustrate a simple model of the author-paper-citation relationship. Both opportunities and limitations have been sketched. The use of a joint model for studies of author productivity and impact at higher levels of aggregation is a topical issue in scientometrics: Hitherto the celebrated but also disputed h-index (Hirsch, 2005), originally proposed for the assessment of research performance at the micro level, was the only one that has combined these two aspects, and afterwards been extended for the use at higher aggregation level in the context of institutional and journal evaluation as well.

For illustration purposes, we have selected authors with ResearcherID and active in 2011 in order to exclude errors in author identification as far as possible. Of course, we have to mention that homonyms and synonyms still occur in RIDs too (cf. Heeffer et al., 2013) but the weight of errors is reasonably small. The main advantage of this model is the possibility of studying citation impact under the condition of the author's productivity, and the identification of high performance in terms of both productivity and impact. However, the same precision as experienced with "classical" citation indicators defined on paper sets could not be reached. The main problem is of conceptual nature: Authors and their papers might hold a different position in various environments created by co-authorship of subject-related issues. This has already induced Hirsch to revise his index in terms of co-authorship (Hirsch, 2010). His new indicator also substantiated that complex constellations cannot be described by separately fractionated parts of the model.

The conclusions drawn from this study are two-fold: On the one hand, author-identification systems need to extended in a reliable way to reach a nearly complete coverage of the author population in the database so that indicators based on author IDs can be considered representative enough to be used as reference standards. The limited discriminative power of author-based indicators and the heterogeneity of the underlying author population, on the other hand, prevents the use of the indicators for the analysis of individual research performance as well as in the context of fine-grained benchmark studies at higher levels of aggregations.

Finally, we would like to emphasise again the necessity and general use of the model introduced in this study, which is formally independent of any author-identification system. The model makes is possible to formalise and describe the relationship between authors, their

publications and the citations those publications receive. The neglect of the structural properties and peculiarities of this "triangle relationship" might result in misinterpretation or even miscalculation of statistics and indicators at this level. The use of author identification in this context is an important means of demonstrating the measurement of this relationship for at least a considerable share of active authors.

References

Braun, T., Glänzel, W., & Schubert, A. (2001), Publication and cooperation patterns of the authors of neuroscience journals. *Scientometrics*, *51*(3), 499–510.

Caron, E. & van Eck, N.J. (2014), *Large scale author name disambiguation using rule-based scoring and clustering.* In: E. Noyons (Ed.), *"Context Counts: Pathways to Master Big and Little Data". Proceedings of the STI Conference 2014, Leiden University, 2014*, 79–86.

Glänzel, W. & Schoepflin, U. (1994), A stochastic model for the ageing analyses of scientific literature. *Scientometrics*, *30*(1), 49–64.

Glänzel, W. & Schoepflin, U. (1995), A bibliometric study on ageing and reception processes of scientific literature. *Journal of Information Science*, *21*(1), 37–53.

Glänzel, W. & Schubert, A. (2003), A new classification scheme of science fields and subfields designed for scientometric evaluation purposes. *Scientometrics*, *56*(3), 357–367.

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., Thijs, B., & Debackere, K. (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.

Glänzel, W., Thijs, B., & Debackere, K. (2014), The application of citation-based performance classes to the disciplinary and multidisciplinary assessment in national comparison and institutional research assessment. *Scientometrics*, *101*(2), 939–952.

Heeffer, S., Thijs, B., & Glänzel, W. (2013), Are registered authors more productive? *ISSI Newsletter*, *9*(2), 29–32.

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.

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*(3), 741–754.

Moed, H.F., van Leeuwen, T.N., & Reedijk, J. (1998), A new classification system to describe the ageing of scientific journals and their impact factors. *Journal of Documentation*, *54*(4), 387–419.

Price, D.D. & Gürsey, S. (1976), Studies in scientometrics. Part 1. Transience and continuance in scientific authorship. *International Forum on Information and Documentation*. *1*, 17–24.

Schubert, A. & Glänzel, W. (1984), A dynamic look at a class of skew distributions. A model with scientometric applications. *Scientometrics*, *6*(3), 149–167.

Schubert, A. (2007), Successive h-indices. *Scientometrics*, *70* (1), 201–205.

Strotman, 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*(9), 1820–1833.

Tang, L. & Walsh, J.P. (2010), Bibliometric fingerprints: name disambiguation based on approximate structure equivalence of cognitive maps. *Scientometrics*, 84(3), 763–784.

Thomson Reuters (2012), Web of Science® Help. Accessible at: http://images.webofknowledge.com/WOKRS58B4/help/WOS/hp_das1.html. Last modified on 09/18/2012, accessed on 28/12/2014.

A Delineating Procedure to Retrieve Relevant Research Areas on Nanocellulose

Douglas H. Milanez[1] and Ed C. M. Noyons[2]

[1] *douglas@nit.ufscar.br*
Federal University of Sao Carlos, Centre for Information Technology in Materials (NIT/Materiais), Washington Luis Highway, km 235, São Carlos – SP (Brazil)

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

Abstract
Advances concerning publication-level classification system have been demonstrated striking results by dealing properly with emergent, complex and interdisciplinary research areas, such as nanotechnology and nanocellulose. However, less attention has been paid to propose a delineation procedure using specific subjects and understand how it could provide interesting regards about it. This study aimed at proposing a delineation procedure to retrieve relevant research areas addressed to nanocellulose using the research areas clustered by the CWTS Web of Science Publication-level Classification System. The procedure involved an iterative process, which includes developing and cleaning set of core publication regarding the subject and analysis of which cluster they might be associated. Nanocellulose was selected as the subject of study. A discussion about each step of the procedure was also provided. The proposed delineation procedure enabled to retrieve relevant publications from research areas involving nanocellulose. Twelve research topics were identified, mapped and associated with current research challenges on nanocellulose.

Conference Topic
Methods and techniques

Introduction
In recent years, bibliometrics has been used often to monitor and quantitatively assess scientific fields within the context of science policy and research management (Moed, Glänzel, & Schmoch, 2004; Okubo, 1997; Raan, 2014). Partly, it is a consequence of the increased use of Internet since the early 1990s and the development of information technologies. Together, they made a huge volume of scientific databases available. Meanwhile, scientific studies have become more complex and interdisciplinary, involving the exchange of knowledge between scientists from different disciplines. Nanotechnology-focused research is a good example. Bibliometric indicators and tools are useful instruments to study and gain insight in science and, in particular, complex fields or research areas, c.f., van Raan (2004). Therefore, many studies on nanotechnology relied on bibliometric approaches (Hullmann & Meyer, 2003; Igami, 2008; Kostoff, Koytcheff, & Lau, 2009; Milanez, Faria, Amaral, Leiva, & Gregolin, 2014; Mogoutov & Kahane, 2007; Wang, Notten, & Surpatean, 2012). The problems often are: how to delineate a field or research area, how to retrieve the relevant data, and which publications to include and which not.

In this sense, classification systems have been used as an indispensable tool to study the structure and dynamics of scientific fields (Boyack, Klavans, & Börner, 2005; Glanzel & Schubert, 2003; Leydesdorff, Carley, & Rafols, 2013; Waltman & van Eck, 2012). They can simplify literature search and retrieving procedures (Glanzel & Schubert, 2003; Waltman & van Eck, 2012). According to Glanzel and Schubert (2003), classification of science into a disciplinary structure can be as old as science and, currently, most of them are based on journal assignment, such as the Web of Science and Scopus systems. The drawback of these

journal-based classification systems is the fact they do not deal properly with multidisciplinary journals or interdisciplinary research (Waltman, van Eck, & Noyons, 2010).

The development of publication-level classification systems has been a current subject of research (Boyack et al., 2011; Waltman & van Eck, 2012). Boyack et al. (2011) clustered a corpus of 2.15 million biomedical publications from Medline database (2004-2008) which generated coherent and concentrated cluster solution of text-based similarity approaches based on keywords extracted from titles and abstracts. They found their approach more precise than the Medical Subject Headings. Waltman and van Eck (2012) proposed a methodology to clustering a large-scale set of scientific publication indexed on Thomson Reuters' Web of Science database. Each publication was assigned to a single research area, which was organized in a three-level hierarchical structure. Their methodology took into account direct citation to cluster the publication. They labelled each research area with discriminative keywords extracted from titles and abstracts. Such publication-level classification systems may be used to gain insights on research areas involved in specific subjects.

In the present study, we intended to map relevant research areas associated with nanocellulose, which is a sustainable nanomaterial that has a great potential for innovation (Isogai, 2013; Mariano, Kissi, & Dufresne, 2014; Milanez, Amaral, Faria, & Gregolin, 2013; Moon, Martini, Nairn, Simonsen, & Youngblood, 2011). Nanocellulose has been a research area for many countries, including the major producers of cellulose worldwide, such as the USA, Canada, Finland, Sweden and Brazil (Milanez et al., 2013). Different disciplines are involved with nanocellulose research since its properties and behaviour have allowed applications as reinforcement agent in composite materials, packing material, optically transparent paper for electronic devices, texturizing agent in cosmetics and food, bio-artificial implants and bandages (Isogai, 2013; Klemm et al., 2011; Mariano et al., 2014; Moon et al., 2011; Siqueira, Bras, & Dufresne, 2010).

Nanocellulose is a generic term referring to cellulose nanofibrils on the one hand and cellulose nanocrystals on the other (Dufresne, 2013; Klemm et al., 2011; Moon et al., 2011; Siqueira et al., 2010; TAPPI, 2011). Cellulose nanocrystals are basically shorter and rod-like crystalline cellulose, whereas cellulose nanofibrils are long chains of alternate amorphous and crystalline cellulose. Consequently, they differ on their mechanical and functional properties (Eichhorn et al., 2010; Mariano et al., 2014; Moon et al., 2011). Both types of nanocellulose can be obtained from renewable sources, including natural fibres, plants, pulp and forest and agricultural residues. Moreover, cellulose nanocrystals can be biosynthesized by bacteria, resulting in the also called bacterial cellulose (Klemm et al., 2011; Milanez et al., 2013; Moon et al., 2011).

Checking the research topics associated with nanocelluloses will provide insights into current technical challenges concerning this nanomaterial, such as increasing the scale of production minimizing costs, characterization of sources and mechanical properties. Surface modifications to reduce moisture adsorption and improve the adhesion between the nanomaterial and the polymeric matrix, thermal degradation, and biocompatibility with living tissues has also been target of research (Gardner, Opo, Oporto, Mills, & Samir, 2008; Isogai, 2013; Klemm et al., 2011; Mariano et al., 2014; Milanez et al., 2013; Moon et al., 2011; Siqueira et al., 2010).

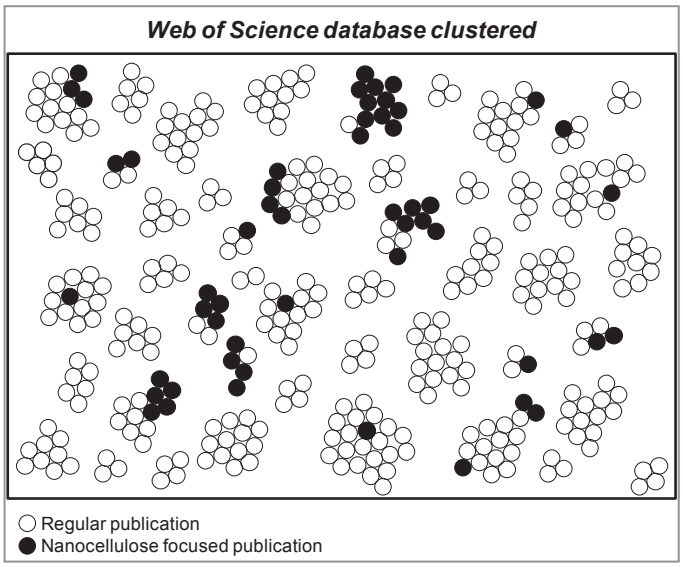
This study aims at proposing a delineation procedure to retrieve relevant research areas addressed to a specific topic. Nanocellulose was selected as a case, but it may be used for other subjects, of course. The approach involves research areas identified in the CWTS Web of Science Publication-level Classification System, a 2014 update of the version introduced by Waltman & van Eck (2012). This paper is structured as follows. In the next section, we describe the overall delineating procedure and its general issues. Next, we discuss details

concerning specific parts and tasks. We present and discuss results in Section 3 and finally in Section 4 we draw our conclusions.

Methodology

Overall delineation procedure

To delineate the field, i.e., to collect a relevant set of publications to represent it, we will select clusters from the CWTS publication level classification system. By this method we will identify papers that will not easily be picked up by keyword or journal based search strategies. Figure 1 presents a schematic representation of the distribution of the clustered Web of Science publications according to CWTS Publication-level classification system (Waltman & van Eck, 2012). Predefined nanocellulose publications are indicated as black circles and the first step is retrieving all research area that contains at least one of them.



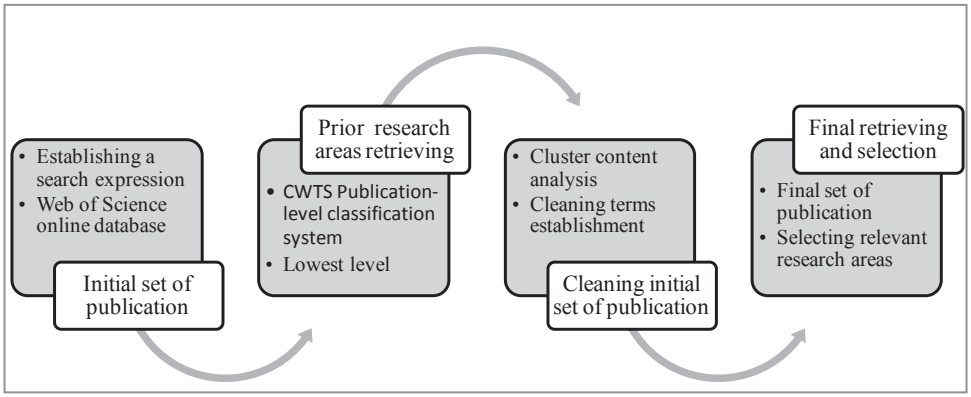
Source: authors.

Figure 1. Schematic representation of Web of Science publications clustered according to the CWTS Publication-level Classification System. The black nodes represent the publications focused on nanocellulose.

Figure 2 depicts the proposed procedure as an iterative process which can be described in four main steps:

1. *Determine an initial set of publication concerning the theme of interest.* In this first step, a set of publication which well represents the theme of interest (nanocellulose) is retrieved via the online Web of Science database, using a straightforward search strategy. This set of publication is a starting set and will be refined as well as expanded through the next steps;
2. *Prior retrieval of nanocellulose research areas*. The second step involves locating the research areas (publication clusters) with at least one publication from the initial set of nanocellulose. The bottom level of the classification scheme was used in this study (Waltman & van Eck, 2012);
3. *Analysis of retrieved research area and cleaning of the initial set*. The content of each research area was analysed pragmatically. A cleaning task was developed by selecting terms to eliminate part of the initial set of nanocellulose publication. This step provided a final set of nanocellulose publication clusters and enhanced the precision of research area assigned to nanocellulose;

4. *Final retrieval and selection of relevant nanocellulose research areas*. After cleaning the initial set of nanocellulose publication, the research areas (publication clusters) were retrieved again. Finally, as the number of topics retrieved was high, a selection that relies on the 80/20 rule was conducted reaching the final research areas associated with nanocelulose.



Source: authors.

Figure 2. Iterative process of the overall procedure proposed.

Determine an initial set of publication on nanocellulose

A search expression was developed considering several terms and synonyms recommended by experts and found in nanocellulose literature (Klemm et al., 2011; Milanez et al., 2013; Siqueira et al., 2010; Siró & Plackett, 2010), as can be seen from Table 1. The search expression encompassed different words that refer to cellulose nanocrystals, cellulose nanofibrils, and bacterial cellulose as well as other generic forms, such as nanocellulose, cellulose nanoparticles, and cellulose nanofiller. The search was conducted in March 31th 2014 in the online Web of Science database (topic search). Only articles that attended the CWTS Web of Science publication-level classification system criteria[1] were used, though.

Table 1. Boolean search expression to retrieve the initial set of nanocellulose publications.

("bacterial cellulos*") OR ("cellulos* crystal*") OR ("cellulos* nanocrystal*") OR ("cellulos* whisker*") OR ("cellulos* microcrystal*") OR ("cellulos* nanowhisker*") OR ("nanocrystal* cellulos*") OR ("cellulos* nano-whisker*") OR ("cellulos* nano-crystal*") OR ("nano-crystal cellulos*") OR ("cellulos* micro-crystal*") OR ("cellulos* microfibril*") OR ("microfibril* cellulos*") OR ("cellulos* nanofibril*") OR ("nanofibril* cellulos*") OR ("micro-fibril* cellulos*") OR ("nano-fibril* cellulos*") OR ("cellulos* micro-fibril*") OR ("cellulos* nano-fibril*") OR ("cellulos* nanofiber*") OR ("nanocellulos*") OR ("cellulos* nanoparticle*") OR ("nano-cellulos*") OR ("nanoparticl* cellulos*") OR ("nanosiz* cellulos*") OR ("cellulos* nanofill*") OR ("nano-siz* cellulos*") OR ("cellulos* nano-fiber*") OR ("cellulos* nano-particle*") OR ("cellulos* nano-fill*") OR ("nano-particl* cellulos*"))

Source: Developed considering nanocellulose-focused terms found in the literature (Klemm et al., 2011; Milanez, Amaral, Faria, & Gregolin, 2013; Siqueira, Bras, & Dufresne, 2010; Siró & Plackett, 2010) and expert opinions.

Prior retrieval of nanocellulose research areas

Research areas that contained at least one publication from the nanocelulose set were retrieved from the CWTS Web of Science Publication-level database. In total, 533 research

[1] The classification system takes into account only *article*, *letter* and *review* published from 2000 to 2013 and indexed in the Science Citation Index Expanded and the Social Science Citation Index. Moreover, to be part of one research area, a publication must be related, either directly or indirectly, to at least 49 other publications in terms of citation (Waltman & van Eck, 2012).

topics were found. These clusters showed large differences in terms of volume (number of publications included). The largest cluster contains 2,751 publications whereas the smallest one covers only 50 publications. Almost 80% of these clusters contained less than three publications from the initial set.

Interestingly, we found that two research areas (clusters) included 56.3% of the initial nanocellulose set of publications. Moreover, in these two clusters, more than 80% overlapped with the initial set. Their descriptive labels also pointed towards nanocellulose research. Therefore, they were considered as nuclei of research in nanocellulose. Other clusters in which the representation of the initial set was much lower, were considered peripheral research areas and their relevance to nanocellulose research was evaluated (see next section).

Analysis of retrieved research area and cleaning of the initial set

An analysis of the content of publications in the peripheral research areas was conducted. We wanted to check whether these articles focused on the nanomaterial as an object of research. If not they were considered noise. Because an evaluation of all research area retrieved would be too labour intensive, we made a selection. The checking task was performed only on those clusters that matched one of the following criteria:

- Research topics that contained at least 20 publications from initial dataset;
- Research topics of which at least 5% overlapped (percentage proportion) with the initial set.

A total of 20 (peripheral) clusters were evaluated. The analysis regarded only articles from the initial dataset. The task involved reading each title to decide whether the article was a study focused upon nanocellulose or not. When the title was not clear, the abstract was also consulted.

Once the checking process was completed, specific terms were identified to clean the initial set of nanocelulose publications. Only research topics with high percentage of "noise publication" were used[2]. Noun-phrases were obtained with support of VOSviewer corpus map analysis applied to titles and abstracts from publications belonging to these clusters. Table 2 present the terms used to clean the nanocellulose-focused publications retrieved using the search expression from Table 1. They were applied on the title, abstract, author's keyword and keyword plus search field. The effect of this cleaning task on the nuclei clusters and the peripheral clusters we used will be discussed in the results.

Table 2. Boolean expression of terms used to clean the nanocellulose-focused publications.

"gene" OR "xyloglucan" OR "microtubule" OR "*cyto*" OR "kinesi" OR "tubulin" OR "*cell wall*" OR "spindle" OR "phragmoplast" OR "mitosis" OR "preprophase" OR "phenotype" OR "*plant growth*" OR "meiosi" OR "*lignin distribution*" OR "delignification" OR "hemicellulose" OR "saccharification" OR "ethanol yield" OR "lignocellulos*" OR "glucosidase" OR "xylanase"

Source: Authors.

Final retrieving and selection of relevant research areas

The final set of nanocellulose publication comprised 2,600 nanocellulose publications (named now as core-nanocellulose) and they were assigned to 428 research areas, which still would be a highly number of cluster to be evaluated. Furthermore, 81.0% of these clusters included only one or two publications from the core-nanocellulose publication, which questions their actual relevance to the advances on nanocelulose studies. Therefore, a selecting step was introduced.

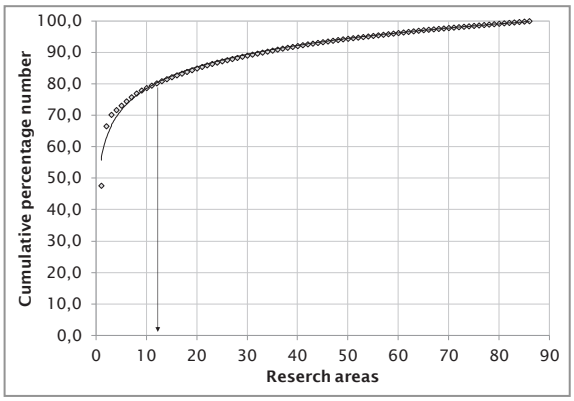
We introduce here the *Pareto Principle* (or 80/20 rule). This principle states that "roughly 80% of the effects come from 20% of the causes" (Juran & Godfrey, 1998) and is found in

[2] The presence of "noise publications" is usual in bibliometric analysis because there is no exhaustive search.

bibliometric and library studies (Gupta, 1989; Kao, 2009; Stephens, Hubbard, Pickett, & Kimball, 2013). We hypothesize that 80% of the core set will be assigned to 20% of the areas. To reach these relevant research areas, the steps below were carried out:

1. The research areas were listed in descending order of the total number of publications from the core-nanocellulose;
2. Research topics with one or two publications from the core-nanocellulose were excluded[3]. This yields 85 research areas remaining;
3. The representativeness of each research area was calculated by the number of publication of the core-nanocellulose of that cluster divided by 2,200 (which is the total of publication found in the 85 remaining research areas);
4. The cumulative percentage number of publications from the core-nanocellulose was obtained summing the values from the step before, as can be seen from Figure 3. The number of research to be assessed was those where the cumulative percentage number of publication reach approximately 80%.

We found that twelve research areas covered the required 80%, which means 14.1% of the total of 85 research topics. We do not claim that our selecting procedure was perfect, but a quick analysis of the chosen research topics showed themes currently found in nanocellulose literature.



Source: CWTS Web of Science Publication-level database.

Figure 3. Cumulative percentage number to research areas with six or more publications from the core-nanocellulose.

Independency test

An independency test was conducted to evaluate the effectiveness of the procedure proposed. The test involved retrieving the number of publication from the top five authors before and after cleaning and selecting the relevant research areas. The percentage decreases of their overall number of publication and from their main cluster were verified.

Results and discussion

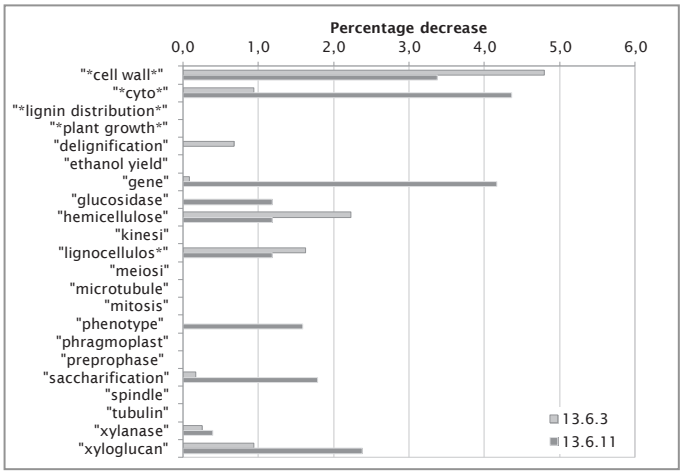
In this section we discuss the effect of cleaning up the core set of publications by using 'cleaning terms', i.e., terms to increase the accuracy of our initial set. Moreover, we present a basic structure of the field on the basis of the delineation we developed.

Effect of cleaning the initial set of nanocellulose publications

Half of the 22 terms we used to clean the nanocellulose search strategy did not affect the coverage of core-nanocellulose publications in the nuclei research areas, as depicted in Figure

[3] According to Waltman and van Eck (2012), the lowest research area contains 50 publications, consequently, clusters with less than 1% of proportion were not accounted for.

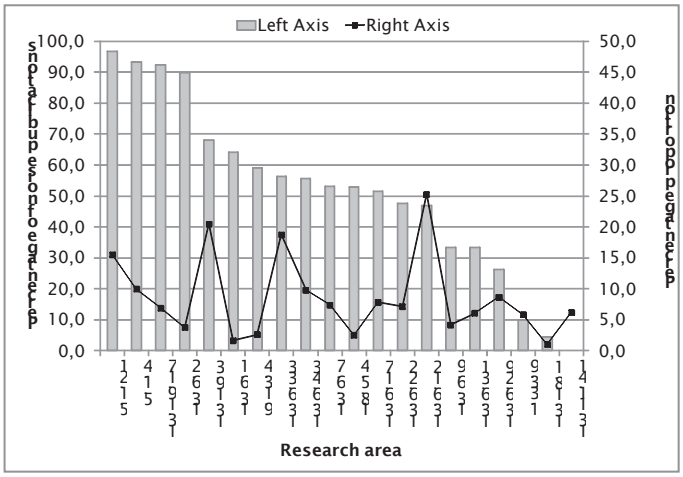
4. To the other half, none term could reduce the coverage in more than 5%. The terms that influenced research area 13.6.4 the most were "*cell wall*" and "hemicelluloses" while "*cyto*", "gene" and "*cell wall*" were the ones that decreased the most core-nanocellulose coverage in cluster 13.6.11. Overall, research topic 13.6.11 had its core-nanocellulose publication reduced in 17.5% while the decrease to cluster 13.6.3 was 10.2%. Nonetheless, both clusters still concentrated publication from the core-nanocellulose after the cleaning tasks (the proportion was 74.0% to research area 13.6.3 and 72.1% to 13.6.11). Therefore, they still had the status of nuclei research areas.



Source: CWTS Web of Science Publication-level database.

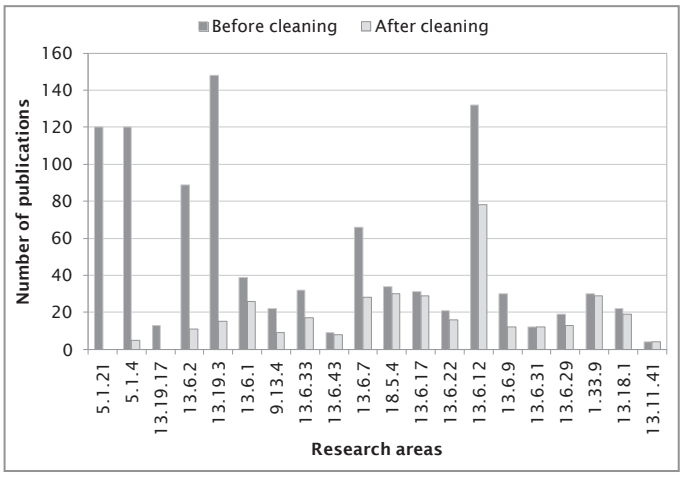
Figure 4. Effect of cleaning terms on the number of publication from nuclei research areas.

As to the 20 peripheral research topics whose nanocellulose set of publication were evaluated, no direct correlation was observed between the proportional relevance of each clusters and the percentage of noise, according to Figure 5. Four research topics had a high percentage (>70%) of 'noisy' publications mainly focusing on biological issues of plants, ethanol production, and enzymes aspects, not having the nanomaterial as a final object of research. Since these four were used to select the cleaning terms, the cleaning affected them highly. Two of them were even eliminated. Furthermore, other peripheral clusters had their nanocellulose publication coverage diminished, as shown on Figure 6.



Source: CWTS Web of Science Publication-level database.

Figure 5. Percentage of noise of core-nanocellulose publications and proportion between core-nanocellulose publications and total number of publications over research area.



Source: CWTS Web of Science Publication-level database.

Figure 6. Effect of cleaning terms on the number of publication from selected peripheral research areas.

Effect of cleaning procedure on top five authors (independency test)

A second test verified the effect of the cleaning process on the coverage of key-authors (top 5). The decrease in the number of publication is presented in Table 3. All authors concentrated their publications on nuclei research topics, mainly on 13.6.3. Only author E focuses primarily on research area 13.6.11. Although the result shows that the overall number of publication diminished in more than 10%, their position as the top authors did not changed but for author E, who went down to the seventh position. It should be noted, however, that research area 13.6.11 was affected more by the cleaning procedure than 13.6.3.

Table 3. Effect of on main authors publications.

Author	*Number of publication*		*Decrease (%)*	
	*Before**	*After**	*Overall*	*Nuclei*
A	87	78	-10,3	-6,33
B	51	40	-21,6	-14,9
C	50	43	-14	0
D	50	39	-22	-18,2
E	48	29	-39,6	-26,5

* Before and after the cleaning step.

Source: CWTS Web of Science Publication-level database.

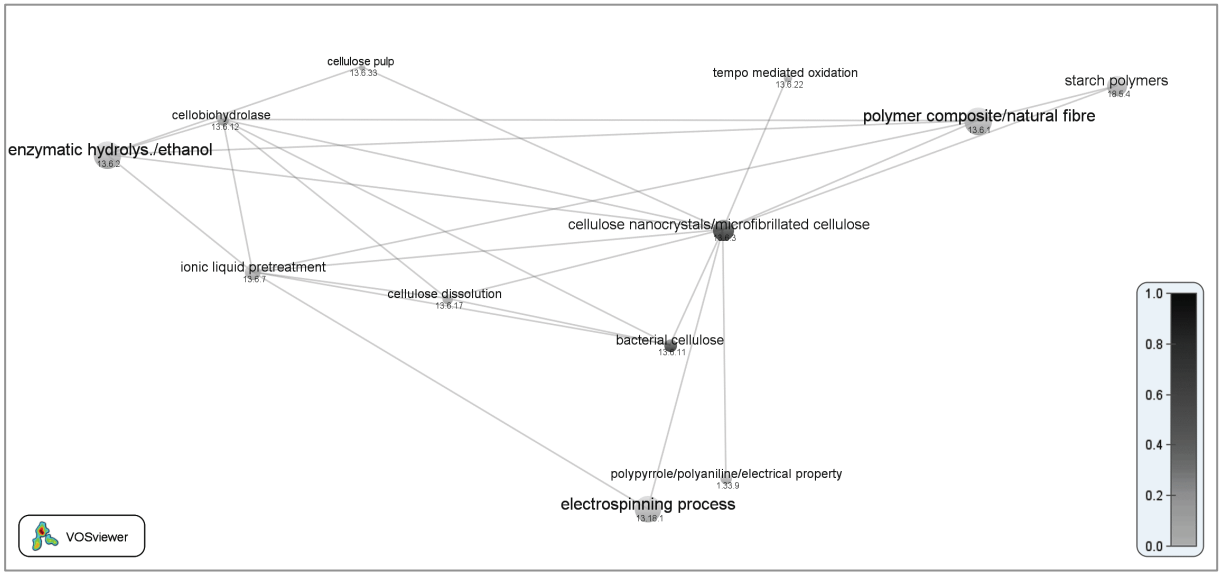
Map of the Nanocellulose research topics

The delineating approach was able to retrieve two nuclei research areas, one associated with cellulose nanocrystals and nanofibrils and other to bacterial cellulose. The peripheral research topics regards biodegradable polysaccharides (starch polymers), polymer composites based on natural fibres, and intrinsically conducting polymers. Other peripheral research areas included enzymatic hydrolyses and ethanol production, cellobiohydrolyse, cellulose pulp and cellulose dissolution, and ionic liquid pre-treatment. Electrospinning process and tempo mediated oxidation, which is an treatment that uses the chemical compound (2,2,6,6-Tetramethylpiperidin-1-yl)oxy (TEMPO), were also part of the final selection. These themes appears frequently in nanocellulose-focused studies (Azizi Samir, Alloin, & Dufresne, 2005; Charreau, Foresti, & Vazquez, 2013; Chirayil, Mathew, & Thomas, 2014; Dai et al., 2014; Domingues, Gomes, & Reis, 2014; Durán, Lemes, & Seabra, 2012; Eichhorn et al., 2010; Isogai, 2013; Klemm et al., 2011; Moon et al., 2011; Orts et al., 2005; Pääkkö et al., 2007; Siqueira et al., 2010; Siró & Plackett, 2010)

Figure 6 presents a map with these research topics (nodes). The map positions the topics on the basis of their citation relations. The closer two topics, the more frequent the citation traffic between them. The node labels match the main content of the clusters. Moreover, all selected clusters had their set of nanocellulose publication evaluated in the cleaning task.

The nuclei research areas are darker and positioned in the centre of the map. Research area 13.6.3 (cellulose nanocrystals/microfibrillated cellulose) has citation connections to all clusters. On the other hand, research topic 13.6.11 is connected only with four other clusters, which might indicate its lower relevance than the other nucleus research area. At the top right of the map are located two research areas addressed to starch polymers and polymer composites based on natural fibres. These research topics regard the development of sustainable materials (Durán et al., 2012; Moon et al., 2011; Siqueira et al., 2010; Isogai, 2013).

Research area concerning enzymatic hydrolysis is highly close to the research topic cellobiohydrolase, i.e., enzymes that perform the process of hydrolyse, and ionic liquid pre-treatment, which also relies on enzymatic approaches. However, they were located further than the nuclei clusters. Indeed, one of them was considered as highly noisy (13.6.2), but we should take into account that nanocellulose obtainment has been also studied as a secondary product of bio-ethanol production (Beecher, 2007; Zhu, Sabo, & Luo, 2011). Moreover, enzymatic pre-treatment has been researched to improve nanocellulose defibrillation (Pääkköet al. 2007; Moon et al., 2011; Klemm et al., 2011; Siqueira et al., 2010; Isogai, 2013).



Source: CWTS Web of Science Publication-level database.

Figure 6. Selected research area according to the procedure proposed.

At the bottom of the map, electrospinning process and conductive polymers were positioned closely, but there is no citation connection between them. Electrospinning is a technique used to produce micro- and nano-sized polymer-based fibres, and nanocellulose has been studied to improve the mechanical property of the final fibre (Dai et al., 2014). Nanocellulose electrical and magnetic properties have also been explored to be used with conductive polymers (Moon et al., 2011; Klemm et al., 2011). The other three research areas (cellulose pulp, cellulose dissolution and tempo mediated oxidation) are the smallest ones, and probably the publications that belong to them might be associated with other clusters on new updates performed using the classification system (Waltman & van Eck, 2012). Tempo mediated

oxidation is a current technique to perform pre-treatment of nanocellulose (Klemm et al., 2011; Isogai, 2013).

Conclusion

The proposed delineation procedure enabled us to retrieve relevant publications from research areas involving nanocellulose. Twelve research topics were identified, mapped and associated with current research challenges on nanocellulose. Two of them were highlighted as nuclei since they contain most part of the initial set of publications. The effect of the cleaning step on nuclei and peripheral clusters provided valuable feedback and demonstrated its importance to establishing relevant clusters afterwards. The independency test showed that the cleaning procedure could have been too rigorous and further research should be carried out to understand how it affected core authors' publication.

Delineating scientific fields is a complex task as boundaries are not frequently well established since scientific studies have become more complex and interdisciplinary. More and more exchange of knowledge between scientists from different disciplines is involved. Our approach retrieves and delineates the real nuclei and the peripheral research areas concerning nanocellulose studies. This clear separation provides suggestions for further research, putting the nuclei research in context. One of the ideas involves the knowledge flow from peripheral research topics to the nuclei areas. We intend to map how they provide the necessary knowledge to face nanocellulose current challenges and how country and scientific institutions are contributing to this evolution.

Acknowledgments

The authors are grateful to the São Paulo Research Foundation (process number 2012/16573-7) and comments from researchers of CWTS and NIT/Materiais. We are also thankful to the Graduate Program in Materials Science and Engineering at the Federal University of São Carlos for supporting this work.

References

Azizi Samir, M. A. S., Alloin, F., & Dufresne, A. (2005). Review of recent research into cellulosic whiskers, their properties and their application in nanocomposite field. *Biomacromolecules*, *6*(2), 612–26. doi:10.1021/bm0493685

Beecher, J. (2007). Wood , trees and nanotechnology. *Nature Nanotechnology*, *2*(August), 466–467.

Boyack, K. W., Klavans, R., & Börner, K. (2005). Mapping the backbone of science. *Scientometrics*, *64*(3), 351–374.

Boyack, K. W., Newman, D., Duhon, R. J., Klavans, R., Patek, M., Biberstine, J. R., … Börner, K. (2011). Clustering more than two million biomedical publications: comparing the accuracies of nine text-based similarity approaches. *PloS One*, *6*(3), e18029. doi:10.1371/journal.pone.0018029

Charreau, H., Foresti, M. L., & Vazquez, A. (2013). Nanocellulose patents trends: a comprehensive review on patents on cellulose nanocrystals, microfibrillated and bacterial cellulose. *Recent Patents on Nanotechnology*, *7*(1), 56–80. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/22747719

Chirayil, C. J., Mathew, L., & Thomas, S. (2014). Review of recent researhc in nanocellulose preparation from different lignocellulosic fibers. *Review of Advanced Materials Science*, *37*, 20–28.

Dai, L., Long, Z., Ren, X., Deng, H., He, H., & Liu, W. (2014). Electrospun polyvinyl alcohol/waterborne polyurethane composite nanofibers involving cellulose nanofibers. *Journal of Applied Polymer*, *41051*, 1–6. doi:10.1002/app.41051

Domingues, R. M. A., Gomes, M. E., & Reis, R. L. (2014). The potential of cellulose nanocrystals in tissue engineering strategies. *Biomacromolecules*, *15*, 2327–2346.

Dufresne, A. (2013). Nanocellulose: A new ageless bionanomaterial. *Materials Today*, *16*(6), 220–227.

Durán, N., Lemes, A. P., & Seabra, A. B. (2012). Review of cellulose nanocrystals patents: preparation, composites and general applications. *Recent Patents on Nanotechnology*, *6*(1), 16–28. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/21875405

Eichhorn, S. J., Dufresne, A., Aranguren, M., Marcovich, N. E., Capadona, J. R., Rowan, S. J., … Peijs, T. (2010). Review: current international research into cellulose nanofibres and nanocomposites. *Journal of Materials Science*, *45*(1), 1–33. doi:10.1007/s10853-009-3874-0

Gardner, D. J., Opo, Oporto, G. S., Mills, R., & Samir, M. A. S. A. (2008). Adhesion and Surface Issues in Cellulose and Nanocellulose. *Journal of Adhesion Science and Technology*, *22*(5-6), 545–567. doi:10.1163/156856108X295509

Glanzel, W., & Schubert, A. (2003). A new classification scheme of science fields and subfields designed for scientometric evaluation purposes. *Scientometrics*, *56*(3), 357–367.

Gupta, D. . (1989). Scientometric study of biochemical literature of Nigeria, 1970-1984: application of Lotka's Law and the 80/20-rule. *Scientometrics*, *15*(3-4), 171–179.

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

Isogai, A. (2013). Wood nanocelluloses: fundamentals and applications as new bio-based nanomaterials. *Journal of Wood Science*, *59*(6), 449–459. doi:10.1007/s10086-013-1365-z

Juran, J. M., & Godfrey, A. B. (Eds.). (1998). *Juran's quality handbook* (5th ed., p. 1730). New York: McGraw-Hill.

Kao, C. (2009). The authorship and internationality of industrial engineering journals. *Scientometrics*, *81*(1), 123–136. doi:10.1007/s11192-009-2093-4

Klemm, D., Kramer, F., Moritz, S., Lindström, T., Ankerfors, M., Gray, D., & Dorris, A. (2011). Nanocelluloses: a new family of nature-based materials. *Angewandte Chemie (International Ed. in English)*, *50*(24), 5438–66. doi:10.1002/anie.201001273

Kostoff, R. N., Koytcheff, R. G., & Lau, C. G. Y. (2009). Seminal Nanotechnology Literature: A Review. *Journal of Nanoscience and Nanotechnology*, *9*(11), 6239–6270. doi:10.1166/jnn.2009.1465

Leydesdorff, L., Carley, S., & Rafols, I. (2013). Global maps of science based on the new Web-of-Science categories. *Scientometrics*, *94*(2), 589–593. doi:10.1007/s11192-012-0784-8

Mariano, M., Kissi, N. El, & Dufresne, A. (2014). Cellulose nanocrystals and related nanocomposites: review of some properties and challenges. *Journal of Polymer Science*, *52*, 791–806. doi:10.1002/polb.23490

Milanez, D. H., Amaral, R. M. Do, Faria, L. I. L. De, & Gregolin, J. A. R. (2013). Assessing nanocellulose developments using science and technology indicators. *Materials Research*, *16*(3), 635–641. doi:10.1590/S1516-14392013005000033

Milanez, D. H., Faria, L. I. L., Amaral, R. M., Leiva, D. R., & Gregolin, J. A. R. (2014). Patents in nanotechnology: an analysis using macro-indicators and forecasting curves. *Scientometrics*. doi:10.1007/s11192-014-1244-4

Moed, H. F., Glänzel, W., & Schmoch, U. (2004). *Handbook of quantitative science and technology research: the use of publication and patent statistics in studies of S&T systems*. (H. F. Moed, W. Glanzel, & U. Schmoch, Eds.) (Kluwer Aca., p. 785). New York: Kluwer Academic Publishers.

Mogoutov, A., & Kahane, B. (2007). Data search strategy for science and technology emergence: A scalable and evolutionary query for nanotechnology tracking. *Research Policy*, *36*(6), 893–903. doi:10.1016/j.respol.2007.02.005

Moon, R. J., Martini, A., Nairn, J., Simonsen, J., & Youngblood, J. (2011). Cellulose nanomaterials review: structure, properties and nanocomposites. *Chemical Society Reviews*, *40*(7), 3941–94. doi:10.1039/c0cs00108b

Okubo, Y. (1997). Bibliometric Indicators and Analysis of Research Systems: Methods and Examples (No. 01). doi:101787/208277770603

Orts, W. J., Shey, J., Imam, S. H., Glenn, G. M., Guttman, M. E., & Revol, J.-F. (2005). Application of Cellulose Microfibrils in Polymer Nanocomposites. *Journal of Polymers and the Environment*, *13*(4), 301–306. doi:10.1007/s10924-005-5514-3

Pääkkö, M. et al. (2007). Enzymatic hydrolysis combined with mechanical shearing and high-pressure homogenization for nanoscale cellulose fibrils and strong gels. *Biomacromolecules*, *8*(6), 1934–41. doi:10.1021/bm061215p

Raan, A. F. J. Van. (2014). Advances in bibliometric analysis : research performance assessment and science mapping. In W. Blockmans, L. Engwall, & D. Weaire (Eds.), *Bibliometrics: Use and abuse in the review of research performance* (Vol. c, pp. 17–28). Portland.

Reuters, T. (2014). Web of Science. Retrieved March 03, 2014, from http://apps.webofknowledge.com/WOS_GeneralSearch_input.do?product=WOS&SID=4CGNLFV3uyF6vP2Mtvi&search_mode=GeneralSearch

Siqueira, G., Bras, J., & Dufresne, A. (2010). Cellulosic bionanocomposites: a review of preparation, properties and applications. *Polymers*, *2*(4), 728–765. doi:10.3390/polym2040728

Siró, I., & Plackett, D. (2010). Microfibrillated cellulose and new nanocomposite materials: a review. *Cellulose*, *17*(3), 459–494. doi:10.1007/s10570-010-9405-y

Stephens, J., Hubbard, D. E., Pickett, C., & Kimball, R. (2013). Citation Behavior of Aerospace Engineering Faculty. *The Journal of Academic Librarianship*, *39*(6), 451–457. doi:10.1016/j.acalib.2013.09.007

TAPPI. (2011). Roadmap for the development of international standards for nanocellulose (p. 36).

Van Raan, A. F. J. (2004). *Science meansuring. In Handbook of quantitative science and technology research: the use of publication and patent statistics in studies of S&T systems* (pp. 19–50). New York: Kluwer Academic Publishers.

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. doi:10.1002/asi.22748

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. doi:10.1016/j.joi.2010.07.002

Wang, L., Notten, A., & Surpatean, A. (2012). Interdisciplinarity of nano research fields: a keyword mining approach. *Scientometrics*, *94*(3), 877–892. doi:10.1007/s11192-012-0856-9

Zhu, J. Y., Sabo, R., & Luo, X. (2011). Integrated production of nano-fibrillated cellulose and cellulosic biofuel (ethanol) by enzymatic fractionation of wood fibers. *Green Chemistry*, *13*(5), 1339. doi:10.1039/c1gc15103g

Sapientia: the Ontology of Multi-dimensional Research Assessment

Cinzia Daraio[1], Maurizio Lenzerini[1], Claudio Leporelli[1], Henk F. Moed[1], Paolo Naggar[2], Andrea Bonaccorsi[3], Alessandro Bartolucci[2]

[1] *daraio@dis.uniroma1.it; lenzerini@dis.uniroma1.it, leporelli@dis.uniroma1.it; henk.moed@uniroma1.it;*
Department of Computer, Control and Management Engineering Antonio Ruberti, Sapienza University of Rome, via Ariosto, 25 00185 Rome (Italy)

[2] *paolo.naggar@gmail.com; alessandro_bartolucci@fastwebnet.it*
Studiare Ltd., Rome (Italy)

[3] *a.bonaccorsi@gmail.com*
DISTEC, University of Pisa, Pisa (Italy)

Abstract
This paper proposes an Ontology-Based Data Management (OBDM) approach to a multi-dimensional research assessment. It is shown that an OBDM approach is able to take into account the recent trends in quantitative studies of Science, Technology and Innovation, including computerization of bibliometrics, multidimensionality of research assessment, altmetrics, and, more generally, the generation of new indicators with higher granularity and cross-referencing specificities according to increasingly demanding policy needs. The main features of *Sapientia* are presented, the Ontology of Multi-dimensional Research Assessment, developed within a project funded by the University of Rome La Sapienza. Illustrative examples are given of its usefulness for the specification of well known as well as recently developed indicators of research assessment.

Conference Topics
Methods and techniques; Indicators; Science policy and research assessment

Introduction: An Ontology-Based-Data-Management Approach to Multi-Dimensional Research Assessment

The quantitative analysis of Science and Technology is becoming a "big data" science, with an increasing level of "computerization", in which large and heterogeneous datasets on various aspects of Science, Technology and Innovation (STI) are combined. Within this framework, optimistic views, supporting "the end of theory" in favour of data-driven science (Kitchin, 2014), have been opposed to more critical positions in favour of theory-driven scientific discoveries (Frické, 2014) while a more balanced view emerged from a critical analysis of the current existing literature (Ekbia et al., 2015), leading the information systems community to further deeply analyse the critical challenges posed by the big data development (Agarwal, 2014). It has been rightly highlighted that "Data are not simply addenda or second-order artifacts; rather, they are the heart of much of the narrative literature, the protean stuff that allows for inference, interpretation, theory building, innovation, and invention" (Cronin, 2013, p. 435). Moreover, the need for accountability of STI activities to sustain their funding in the current difficult economic and financial situation is increasingly asking for rigorous empirical evidence to support informed policy making. Indeed, the needs to overcome the logic of rankings and the new trends in indicators development, including granularity and cross-referencing, can be explored and exploited in open data platforms with a clear description of the main concepts of the domain (Daraio & Bonaccorsi, 2015). The multidimensionality of research assessment and scholarly impact (Moed & Halevi, 2015), and the recent altmetrics movements (Cronin & Sugimoto, 2014), are questioning the traditional approach in indicators development.

Research assessment, indeed, is becoming increasingly complex due to its multi-dimensionality nature. A Report published in 2010 by the Expert Group on the Assessment of University-Based Research, installed by the European Commission proposed "a consolidated multidimensional methodological approach addressing the various user needs, interests and purposes, and identifying data and indicator requirements" (AUBR, 2010, p. 10). A key notion holds that "indicators designed to meet a particular objective or inform one target group may not be adequate for other purposes or target groups". Diverse institutional missions, and different policy environments and objectives require different assessment processes and indicators. In addition, the range of people and organizations requiring information about university-based research is growing. Each group has specific but also overlapping requirements (AUBR, 2010, p. 51).

Table 1. Main types of research outputs.

Printed outputs (texts)	**Non-printed outputs (non-text)**	**Main type of impact**
Scientific journal paper; book chapter; scholarly monograph	Research data file; video of experiment; software	Scientific-scholarly
Patent; commissioned research report;	New product or process; material; device; design; image; spin off	Economic or technological
Professional guidelines; newspaper article; communication submitted to social media, including blogs, tweets.	Interview; event; art performance; exhibit; artwork; scientific-scholarly advise;	Social or cultural

A research assessment has to take into account a range of different types of research output and impact. As regards output forms, one important distinction is between text-based and non-text based output forms. The main types are presented in Table 1. This table is not fully comprehensive. The specifications of the Panel Criteria in the Research Excellence Framework in the UK (REF, 2012, page 51 a.f.) provide more detailed lists of possible output forms arranged by major research discipline. Table 1 includes forms that are becoming increasingly important such as research data files, and communications submitted to social media and scholarly blogs. A framework for the assessment of these forms is being developed in the field of altmetrics (e.g., Taylor, 2013). The last column indicates the main types of impact a particular output may have. A distinction is made between scientific-scholarly impact, and wider impact outside the domain of science and scholarship, denoted as "societal", a concept that embraces technological, economic, social and cultural impact. A comprehensive overview of the types of impact, and the most frequently used impact indicators is presented in Table 2. The reader is referred to AUBR (2010 and Moed & Halevi (2015) for a further discussion of this table.

It is also important to include the inputs in the analysis; they should be jointly analysed with the outputs to assess the overall impact of the process (see e.g. Daraio et al., 2014, for a conditional multidimensional approach to rank higher education institutions). To meet all these new trends and policy needs a shift in the paradigm of the data integration for research assessment is needed. In this paper we advocate an OBDM approach to research assessment. This new approach radically changes the traditional paradigm of construction of STI indicators and offers a flexible and powerful tool for designing new indicators and develop rigorous policy making. The confidence in this new approach comes from three directions: (i) recent efforts from policy makers to support the creation of new datasets on S&T; (ii) bottom up standardization initiatives; (iii) development of almetrics and web-based indicators. To start with, in the last few years, several initiatives at European level have been based on an intense production and use of new data.

Table 2. Types of Research Impact and Indicators.

Type of impact	Short Description; Typical examples	Indicators (examples)
Scientific-scholarly or academic		
Knowledge growth	Contribution to scientific-scholarly progress: creation of new scientific knowledge	Indicators based on publications and citations in peer-reviewed journals and books
Research networks	Integration in (inter)national scientific-scholarly networks and research teams	(inter)national collaborations including co-authorships; participation in emerging topics
Publication outlets	Effectiveness of publication strategies; visibility and quality of used publication outlets	Journal impact factors and other journal metrics; diversity of used outlets;
Societal		
Social	Stimulating new approaches to social issues; informing public debate and improve policy-making; informing practitioners and improving professional practices; providing external users with useful knowledge; Improving people's health and quality of life; Improvements in environment and lifestyle;	▪ Citations in medical guidelines or policy documents to research articles ▪ Funding received from end-users ▪ End-user esteem (e.g., appointments in (inter)national organizations, advisory committees) ▪ Juried selection of artworks for exhibitions ▪ Mentions of research work in social media
Technological	Creation of new technologies (products and services) or enhancement of existing ones based on scientific research	Citations in patents to the scientific literature (journal articles)
Economic	Improved productivity; adding to economic growth and wealth creation; enhancing the skills base; increased innovation capability and global competitiveness; uptake of recycling techniques;	▪ Revenues created from the commercialization of research generated intellectual property (IP) ▪ Number patents, licenses, spin-offs ▪ Number of PhD and equivalent research doctorates ▪ Employability of PhD graduates
Cultural	Supporting greater understanding of where we have come from, and who and what we are; bringing new ideas and new modes of experience to the nation.	▪ Media (e.g. TV) performances ▪ Essays on scientific achievements in newspapers and weeklies ▪ Mentions of research work in social media

Legend to Table 2: Partly based on AUBR (2010) and Moed & Halevi (2015)

In the field of data on universities, the pioneering efforts of Aquameth (Daraio et al., 2011; Bonaccorsi & Daraio, 2007) and subsequently of Eumida (Bonaccorsi, 2014) have been transformed in an institutional initiative called ETER (European Tertiary Education Register), which will make publicly available microdata on universities in 2015. In the same field, the mapping of diversity of European institutions (Huisman, Meek & Wood, 2007; van Vught, 2009) led to the experimental project U-Map, after which there has been an institutional effort towards a multidimensional ranking exercise, called U-Multiranking (van Vught & Westerheijden, 2010). In the field of Public Research Organisations, there has been an effort to build up a comprehensive list of institutions and to survey their activities within the European Research Area (ERA) context. The results of the large ERA surveys, run in 2013 and 2014, will be made available in 2015. These efforts from Europe have a major counterpart on the other side of the Atlantic, where the STAR Metrics initiative (see https://www.starmetrics.nih.gov/) has promoted a federal and research institution collaboration to create a repository of data and tools that is producing extremely interesting results. All these efforts, however, are based on the construction of new datasets, or the integration of existing datasets into new ones. They do not solve the issue of comparability

and standardization of information and of inter-operability, updating and scalability of databases. It is interesting to observe that, in parallel to these efforts put in place by public institutions and policy makers, there have also been massive bottom up efforts aimed at standardizing the elementary pieces of information. Moreover, these efforts have been based on the construction of partial ontologies. Consider the following.

- ORCID (http://orcid.org/) is a non-profit organization, supported by research organizations, agencies, providers of publication management systems, and publishers, aiming at giving all researchers a unique identifier (ORCID_id number) and keeping it persistent over time. Established at the end of 2009, but operational since end 2012, it has almost reached one million researchers worldwide. Most of the increase has been achieved in a very short time frame: from 100,000 in March 2013 to almost 970,000 as of October 2014 (with 35% from European, Middle East and Asian countries);
- CERIF is a Europe-based initiative aiming at standardizing the operations of funding agencies, with the help of a full-scale ontology of almost all research products (http://www.eurocris.org);
- CASRAI (www.casrai.org) is a Canada-US initiative for the standardization of data on research institutions and funders (also supported by a committee of Science Europe; http://www.scienceeurope.org/scientific-committees/Life-sciences/life-sciences-committee);
- ISNI (www.isni.org) provides lists and metadata on higher education, research, funding and many other types of organizations, while Ringgold (www.ringgold.com) does the same in the world of publishers and intermediaries.

These initiatives are strongly supported by international scientific associations (see for example CODATA http://www.codata.org and the VIVO network of scientists: http://www.vivoweb.org/).

Finally, the rapid growth of alternative metrics and web-based metrics has also created a large space for the production of data from publicly available and other sources (Cronin & Sugimoto, 2014). Summing up, there are powerful trends that point to the need to change the overall philosophy of the production of S&T indicators. Instead of an environment in which indicators are produced in close circles, by constructing ad hoc databases, with no built-in interoperability, updating and scalability features, we have to move towards an environment in which elementary pieces of information are fully standardized, micro-data consistent with standardized definitions are (mostly) publicly available, and indicators are constructed following the policy demands on the basis of stable platforms constantly integrated and updated, instead of starting from scratch each time a new indicator is needed.

Main advantages of an OBDM approach compared to conventional data-base integration approaches

While the amount of data stored in current information systems and the processes making use of such data continuously grow, turning these data into information, and governing both data and processes are still tremendously challenging tasks for Information Technology. The problem is complicated due to the proliferation of data sources and services both within a single organization, and in cooperating environments. The following factors explain why such a proliferation constitutes a major problem with respect to the goal of carrying out effective data governance tasks:

- Although the initial design of a collection of data sources and services might be adequate, corrective maintenance actions tend to re-shape them into a form that often diverges from the original conceptual structure.
- It is common practice to change a data source (e.g., a database) so as to adapt it both to specific application-dependent needs, and to new requirements. The result is that

data sources often become data structures coupled to a specific application (or, a class of applications), rather than application-independent databases.
- The data stored in different sources and the processes operating over them tend to be redundant, and mutually inconsistent, mainly because of the lack of central, coherent and unified coordination of data management tasks.

The result is that information systems of medium and large organizations are typically structured according to a "sylos"-based architecture, constituted by several, independent, and distributed data sources, each one serving a specific application. This poses great difficulties with respect to the goal of accessing data in a unified and coherent way. Analogously, processes relevant to the organizations are often hidden in software applications, and a formal, up-to-date description of what they do on the data and how they are related with other processes is often missing. The introduction of service-oriented architectures is not a solution to this problem per se, because the fact that data and processes are packed into services is not sufficient for making the meaning of data and processes explicit. Indeed, services become other artifacts to document and maintain, adding complexity to the governance problem. Analogously, data warehousing techniques and the separation they advocate between the management of data for the operation level, and data for the decision level, do not provide solutions to this challenge. On the contrary, they also add complexity to the system, by replicating data in different layers of the system, and introducing synchronization processes across layers. All the above observations show that a unified access to data and an effective governance of processes and services are extremely difficult goals to achieve in modern information systems. Yet, both are crucial objectives for getting useful information out of the information system, as well as for taking decisions based on them. This explains why organizations spend a great deal of time and money for the understanding, the governance, the curation, and the integration of data stored in different sources, and of the processes/services that operate on them, and why this problem is often cited as a key and costly Information Technology challenge faced by medium and large organizations today (Bernstein & Haas, 2008).

We argue that ontology-based data management (OBDM, Lenzerini, 2011) is a promising direction for addressing the above challenges. The key idea of OBDM is to resort to a three-level architecture, constituted by the ontology, the sources, and the mapping between the two. The ontology is a conceptual, formal description of the domain of interest to a given organization (or, a community of users), expressed in terms of relevant concepts, attributes of concepts, relationships between concepts, and logical assertions characterizing the domain knowledge. The data sources are the repositories accessible by the organization where data concerning the domain are stored. In the general case, such repositories are numerous, heterogeneous, each one managed and maintained independently from the others. The mapping is a precise specification of the correspondence between the data contained in the data sources and the elements of the ontology.

The main purpose of an OBDM system is to allow information consumers to query the data using the elements in the ontology as predicates. In this sense, OBDM can be seen as a form of information integration, where the usual global schema is replaced by the conceptual model of the application domain, formulated as an ontology expressed in a logic-based language. With this approach, the integrated view that the system provides to information consumers is not merely a data structure accommodating the various data at the sources, but a semantically rich description of the relevant concepts in the domain of interest, as well as the relationships between such concepts. The distinction between the ontology and the data sources reflects the separation between the conceptual level, the one presented to the client, and the logical/physical level of the information system, the one stored in the sources, with the

mapping acting as the reconciling structure between the two levels. This separation brings several potential advantages:

- The ontology layer in the architecture is the obvious mean for pursuing a declarative approach to information integration, and, more generally, to data governance. By making the representation of the domain explicit, we gain re-usability of the acquired knowledge, which is not achieved when the global schema is simply a unified description of the underlying data sources.
- The mapping layer explicitly specifies the relationships between the domain concepts on the one hand and the data sources on the other hand. Such a mapping is not only used for the operation of the information system, but also for documentation purposes. The importance of this aspect clearly emerges when looking at large organisations where the information about data is widespread into separate pieces of documentation that are often difficult to access and rarely conforming to common standards. The ontology and the corresponding mappings to the data sources provide a common ground for the documentation of all the data in the organisation, with obvious advantages for the governance and the management of the information system.
- A third advantage has to do with the extensibility of the system. One criticism that is often raised to data integration is that it requires merging and integrating the source data in advance, and this merging process can be very costly. However, the ontology-based approach we advocate does not impose to fully integrate the data sources at once. Rather, after building even a rough skeleton of the domain model, one can incrementally add new data sources or new elements therein, when they become available, or when needed, thus amortising the cost of integration. Therefore, the overall design can be regarded as the incremental process of understanding and representing the domain, the available data sources, and the relationships between them. The goal is to support the evolution of both the ontology and the mappings in such a way that the system continues to operate while evolving, along the lines of "pay-as-you-go" data integration pursed in the research on data-spaces (Sarma et al., 2008).

The notions of ODBM were introduced in (Calvanese et al. 2007; Poggi et al. 2008), and originated from several disciplines, in particular, Information Integration, Knowledge Representation and Reasoning, and Incomplete and Deductive Databases. The central notion of OBDM is therefore the ontology, and reasoning over the ontology is at the basis of all the tasks that an OBDM system has to carry out. In particular, the axioms of the ontology allow one to derive new facts from the source data, and these inferred facts greatly influence the set of answers that the system should compute during query processing. In the last decades, research on ontology languages and ontology inferencing has been very active in the area of Knowledge Representation and Reasoning. Description Logics (DLs, Baader et al., 2007) are widely recognized as appropriate logics for expressing ontologies, and are at the basis of the W3C standard ontology language OWL. These logics permit the specification of a domain by providing the definition of classes and by structuring the knowledge about the classes using a rich set of logical operators. They are decidable fragments of mathematical logic, resulting from extensive investigations on the trade-off between expressive power of Knowledge Representation languages, and computational complexity of reasoning tasks. Indeed, the constructs appearing in the DLs used in OBDI are carefully chosen taking into account such a trade-off (Calvanese et al., 2007).

As indicated above, the axioms in the ontology can be seen as semantic rules that are used to complete the knowledge given by the raw facts determined by the data in the sources. In this sense, the source data of an OBDI system can be seen as an incomplete database, and query answering can be seen as the process of computing the answers logically deriving from the

combination of such incomplete knowledge and the ontology axioms. Therefore, at least conceptually, there is a connection between OBDM and the two areas of incomplete information (Imielinski & Lipski, 1984) and deductive databases (Ceri et al., 1990).

Sapientia at a glance

The main objective of *Sapientia* is to model all the activities relevant for the evaluation of research and for assessing its impact. For impact, in a broad sense, we mean any effect, change or benefit, to the economy, society, culture, public policy or services, health, the environment or quality of life, beyond academia.

Table 3. Modules of the Sapientia Ontology.

N.	Module Name	Module Description
1	Overview	presents the terminological inventory needed to define the ontology domain: what is to be known to assess research activities and their impact on human knowledge and the economic system
2	Agent	models the individuals involved in the world of research, carrying out knowledge-related activities
3	Activity	models the main knowledge related activities matching them with public and relevant commitments of the agents involved in the domain (each module from 4 to 11 is devoted to a kind of knowledge-related activity - the module name corresponds to the appropriate specialization of the concept *Activity*)
4	*Research activity*	models, among the knowledge-related activities, those that allow the scientific community to advance the state of the art of knowledge
5	*Educational_activity*	models, among the knowledge-related activities, those that allow people to improve their knowledge
6	*Conferring degrees activity*	models, among the knowledge-related activities, those that grant degrees allowing people to widely qualify themselves
7	*Publishing activity*	models, among the knowledge-related activities, those that allow people to know the results of research activities
8	*Preservation activity*	models, among knowledge-related activities, those that permit the preservation of the value of things (related to research activities)
9	*Funding activity*	models, among the knowledge-related activities, those that assign and distribute the funds needed to carry out research, educational and service activities
10	*Inspecting activity*	models, among the knowledge-related activities, those that control and assess research, educational and service activities
11	*Producing activity*	models, among the knowledge-related activities, those that produce economic, society and cultural value
12	Space	models the space and its roles
13	Taxonomy	models the relevant taxonomies that classify the elements of the domain
14	Time	models the depth of time of the domain (this module is spread through the others)

Hence, *Sapientia* covers what is to be known about assess research activities and their impact on human knowledge and the economic system. For this purpose the ontology embraces:

- the inter-relationships between research activities (Modules Research_activity, Publishing_activity);

• the relationships between research activities and people's personal knowledge (Modules Teaching_activity, Conferring_degrees_activity, Publishing_activity, Producing_activity);
• the relationships between research activities and other missions of individuals and institutions (Modules Inspect_activity, Producing_activity);
• the relationship between research activities and the knowledge locally available to the companies in the economic system, enabling their innovative behavior (Module Producing_activity).

The *Sapientia* ontology includes also the activities that are needed for fostering these relationships (Modules Preservation_activity, Inspecting_activity and Funding_activities). The 14 modules that compose *Sapientia* are listed in Table 3.

Modelling choices

We pursued a modelling approach based on processes, which were conceived as collections of activities. A process is composed by inputs and outputs. Individuals and activities are the main pillars of the ontology.

We consider the building of descriptive, interpretative, and policy models of our domain as a distinct step with respect to the building of the domain ontology. However, the ontology will intermediate the use of data in the modelling step, and should be rich enough to allow the analyst the freedom to define any model she considers useful to pursue her analytic goal.

Obviously, the actual availability of relevant data will constrain both the mapping of data sources on the ontology, and the actual computation of model variables and indicators of the conceptual model. However, the analyst should not refrain from proposing the models that she considers the best suited for her purposes, and to express, using the ontology, the quality requirements, the logical, and the functional specification for her ideal model variables and indicators. This approach has many merits, and in particular:

- it allows the use of a common and stable ontology as a platform for different models;
- it addresses the efforts to enrich data sources, and verify their quality;
- it makes transparent and traceable the process of approximation of variables and models when the available data are less than ideal;
- it makes use of every source at the best level of aggregation, usually the atomic one.

More generally, this approach is consistent with the effort of avoiding "the harm caused by the blind symbolism that generally characterizes a hasty mathematization" put forward by Georgescu Roegen in his seminal work on production models and on methods in economic science (Georgescu-Roegen, 1970, 1971, 1979). In fact, one can verify the logical consistency of the ontology and compute answers to unambiguous logical queries.

Moreover, the proposed ontology allows us to follow the Georgescu-Roegen approach also in the use of the concept of process. We can analyze the knowledge production activities, at an atomic level, considering their *time* dimension and such *funds* as the cumulated results of previous research activities, both those available in relevant publications, and those embodied in the authors' competences and potential, the infrastructure assets, and the time devoted by the group of authors to current research projects. Similarly, we can analyze the output of teaching activities, considering the joint effect of *funds* such as the competence of teachers, the skills and the initial education of students, and educational infrastructures and resources. Thirdly, service activities of research and teaching institutions provide infrastructural and knowledge assets that act as a *fund* in the assessment of the impact of those institutions on the innovation of the economic system. The perimeter of our domain should allow us to consider the different channels of transmission of that impact: mobility of researchers, career of alumni, applied research contracts, joint use of infrastructures, and so on. In this context,

different theories and models of the system of knowledge production could be developed and tested (Etzkowitz & Leydesdorff, 2000).

Table 4. Indicators considered for the test of the completeness of Sapientia.

#	Indicator (I)	*Sapientia's Modules*												
		2	3	4	5	6	7	8	9	10	11	12	13	14
I1	Number of published articles	A					F					F,D1	D1	D2
I2	Number of citations	A					F					F,D1	D1	D2
I3	Citations per article	A					F					F,D1	D1	D2
I4	Normalized citation rate	A					F					F,D1	D1	D2
I5	Highly cited publications	A					F					D1	D1	D2
I6	Journal Impact Factor	A					F						F	D2
I7	Subject Normalized Impact Factor	A					F						F	D2
I8	Scimago Journal Ranking Impact Factor	A					F						F	D2
I9	H-index	F					F					A	F	D2
I10	E-index	F					F					A	F	D2
I11	Number of patents	A									F	F,D1	D1	D2
I12	Full text article download count	A					F					F,D1	D1	D2
I13	Mentions in social media	A					F					F,D1	D1	D2
I14	Research output per academic staff	A					F					F,D1	D1	D2
I15	Percentage of Highly Cited Publications	A					F					D1	D1	D2
I16	Number of keynote addresses at conferences	A					F					F,D1	D1	D2
I17	Number of prestigious awards and prizes	A								F		F,D1	D1	D2
I18	Number of visiting research appointments	F,A										D1	D1	D2
I19	Member of editorial board	A					F					D1	D1	D2
I20	Refereeing activity for journals	A	F							F		D1	D1	D2
I21	External research income	A							F		F	D1	D1	D2
I22	Number of competitive grants won	A							F		F	D1	D1	D2
I23	Percentage of competitive grants	A							F		F	D1	D1	D2
I24	External research income per academic staff	A							F		F	F,D1	D1	D2
I25	Employability of PhD graduates	A				F						D1	F,D1	D2
I26	Commerc. of research generated intellectual property	A									F	D1	D1	D2
I27	End-users esteem	A								F		D1	D1	D2
I28	Number of funding from end-users	A,D1							F			D1	D1	D2
I29	Percentage of funding from end-users	A,D1							F			D1	D1	D2
I30	Post-graduate research student load	A		F	F							D1	D1	D2
I31	Involvement of early career researchers in teams	A,F		F								D1	D1	D2
I32	Number of collaborations and partnerships	F		A								F,D1	D1	D2
I33	Doctoral completions	A				F						D1	D1	D2
I34	Research active academics	F,A					F					D1	D1	F,D2
I35	Percentage of research active per total academic staff	F,A					F					D1	D1	F,D2
I36	Total R&D investment	A							F			D1	D1	D2
I37	Research Infrastructures and facilities	A					F	F				D1	D1	D2
I38	Research ethics						F					A,F	F,D1	D2

Testing the Ontology: analysis of the competency questions

One way to check if the ontology contains all the relevant information and/or details to represent the domain of interest, currently used in knowledge representation, is based on the specification of competency questions (Gruninger & Fox 1995). These questions correspond to check whether the ontology contains enough information to answer these types of questions or whether the answers require a particular level of detail or representation of a particular module of the ontology that needs to be further developed. The analysis of the competency questions of *Sapientia* has been carried out on the indicators contained in the paper by Moed and Halevi (2015), integrated with the additional indicators reported in the AUBR (2010) document. In addition, other key references of the ontological commitments have been Moed, Glanzel and Schmock (2004), Moed (2005) and Cronin and Sugimoto (2014), together with the knowledge background of the team of the project.

Table 4 contains the list of indicators considered for the verification of the competency questions. Associated to each indicator are reported the following pieces of information:

- Facts (F) are the content of the data, the relevant information about atomic events relevant for the construction of the indicator;
- Aggregation level (A) is the minimal aggregation level: the concept which classifies the objects included in the indicator;
- Dimensions of the analysis (D), are descriptive properties which are relevant to access higher level of aggregation. They are evaluated by the dimension of taxonomy (D1) and that of time (D2).

Table 5 summarizes the number of facts (F), aggregations (A) and dimensions (D) by module, as reported in Table 4, to check the comprehensiveness of *Sapientia* with respect to the indicators listed therein. Put it in another way, we checked whether our ontology was able to include all the relevant conceptual information requested by the specification of the listed indicators in Table 4. The answer to this question is indeed positive.

Table 5. Some statistics on the "usage" of the Ontology modules.

	2	3	4	5	6	7	8	9	10	11	12	13	14
F	7	1	2	1	2	20	1	7	3	6	13	5	2
A	34	0	1	0	0	0	0	0	0	0	2	1	0
D	2	0	0	0	0	0	0	0	0	0	31	31	38

By inspecting Table 5 it clearly appears that only a few modules are used for the specification of the indicators reported in Table 4. This means that our ontology covers a much broader conceptual domain with respect to the one underlying (even if not formally specified) by the indicators reported in Table 4. The most frequently used module is the Publishing module (7), followed by Space (12) and Funding (9). We note that the modules 12 (Space), 13 (Taxonomy) and 14 (Time) are used in the majority of the cases to further characterize the dimensions of the considered indicators.

A new way to conceive and specify STI indicators

By adopting an OBDM perspective a new approach to designing indicators can be implemented. This new approach aligns very well with the recent trends described in the introduction.

The traditional approach to indicators' design is based on informal definitions expressed in a natural language (English, typically). An indicator is defined as a relationship between variables, e.g. a ratio between number of publications per academic staff, chosen among a predefined set of data collected and aggregated ad hoc, by a private or a public entity, according to the user needs, and hence not re-usable for future assessment and use.

The OBDM approach we pursue in this paper permits a *more advanced specification* of an indicator according to the following dimensions:

- the *ontological dimension*. It represents the domain (portion) of the reality to be measured by the indicator (obviously, in the scope of this paper, all indicators will share the *Sapientia* ontology as their ontological part);
- the *logical dimension*. It denotes the question that has to be asked to the ontological portion in order to retrieve all the information (data) needed for calculating the indicator value. In this case the data are extracted from the sources through the mapping considering the logical specification of the query;
- the *functional dimension*. It indicates the mathematical expression that has to be applied on the result of the logical extraction of data carried out in the previous point in order to calculate the indicator value;
- the *qualitative dimension*. It specifies the questions that have to be asked to the ontological part in order to generate the list of problems affecting the meaningfulness of the calculated indicator. An indicator will be considered meaningful if the list of its problems is empty.

In addition to the advantages of the OBDM recalled in previous sections above, the main specific benefits of this approach for designing indicators are the following:

1. It offers a space to *freely* explore the *generation of new indicators,* not previously specified by users, thanks to the *multiple inheritance* in the hierarchy of the concepts (a concept can be subsumed in several concepts).
2. For standard indicators specified by the users it can be seen immediately what is *missing* or which *problems* exist to calculate them;
3. It provides more alternatives and diagnostic ways to check the *robustness* of indicators with respect to opportunistic behaviour and the general goals of the assessment;
4. The formal specification of the indicators is made *independently* of the data. In this way, when applied to heterogeneous data sources, OBDM offers the opportunity to compute "comparable" indicator values at different level of aggregation. Moreover, it offers a reference system to *check the comparability* level among the heterogeneous sources of data and to identify where to invest in order to overcome the remaining existing comparability problems.
5. This approach permits an *unambiguous* way to define and compute the indicators. The indicator is calculated always in the same way.

Conclusions and further developments

In this paper we advocated the use of an OBDM approach to research assessment. We explained the reasons why a paradigm shift in research assessment is needed and outlined the main advantages of an OBDM approach over traditional databases integration approaches. We described the main objectives and structure of *Sapientia* the Ontology of Multi-dimensional Research Assessment. Finally, we illustrate the new indicator design methodology implicitly provided by an OBDM approach.

Sapientia 1.0 has been closed on the 22nd December 2014 and consisted of around 350 symbols (including concepts, relations and attributes). The full documentation of the Ontology is under way together with the mapping with several sources of data. Due to the works on the documentation and the mapping with the data in progress, as well as the limited number of pages available, we concentrated our presentation on the methodological aspects related to the development of the Sapientia.

We believe in fact that it will open a new stream of studies to further explore and exploit the OBDM approach for STI indicator designers and policy makers.

Acknowledgments

The financial support of the "Progetto di Ateneo 2013", University of Rome La Sapienza, is gratefully acknowledged.

References

Agarwal, R., & Dhar, V. (2014). Editorial—Big Data, Data Science, and Analytics: The Opportunity and Challenge for IS Research. I*nformation Systems Research, 25*(3), 443-448.

AUBR Expert Group (2010). Expert Group on the Assessment of University-Based Research. Assessing Europe's University-Based Research. European Commission – DG Research. EUR 24187 EN

Baader F., D. Calvanese, D. McGuinness, D. Nardi, & P. F. Patel-Schneider, (eds) (2007). *The Description Logic Handbook: Theory, Implementation and Applications.* Cambridge University Press, 2nd edition.

Bernstein P. A. & Haas L.(2008). Information integration in the enterprise. *Communication of the ACM*, *51*(9), 72–79.

Bonaccorsi A., & Daraio C. (eds.) (2007) *Universities and strategic knowledge creation. Specialization and performance in Europe*. Cheltenham, Edward Elgar.

Bonaccorsi, A. (ed.) (2014) *Knowledge, diversity and performance in European higher education*. Cheltenham, Edward Elgar.

Calvanese D., G. De Giacomo, D. Lembo, M. Lenzerini, & R. Rosati (2007), Tractable reasoning and efficient query answering in description logics: The DL-Lite family. *Journal of Automated Reasoning*, *39*(3), 385–429.

Ceri S., G. Gottlob, & L. Tanca (1990). *Logic Programming and Databases*. Springer, Berlin (Germany).

Console M., Lembo D., Santarelli V. & Savo D.F. (2014a). Graphol: Ontology Representation Through Diagrams. *Proc. of the 27th Int. Workshop on Description Logic*.

Console M., Lembo D., Santarelli V., & Savo D.F. (2014b). Graphical Representation of OWL 2 Ontologies through Graphol. *Proc. of the 13th International Semantic Web Conference Posters & Demos*.

Cronin B. & Sugimoto C. (ed) (2014). Beyond bibliometrics. Harnessing multidimensional indicators of scholarly impact. MIT Press, Cambridge Mass.

Cronin, B. (2013). Thinking about data. *Journal of the American Society for Information Science and Technology*, *64*(3), 435–436.

Daraio, C. et al. (2011). The European university landscape: A micro characterization based on evidence from the Aquameth project. *Research Policy 40*, 148–164.

Daraio C. & Bonaccorsi A. (2015). Beyond university rankings? Generating new indicators on universities by linking data in open platforms. under review for *JASIST*.

Daraio, C., Bonaccorsi A., & Simar L. (2015). Rankings and university performance: a conditional multidimensional approach. *European Journal of Operational Research*, *244*, 918–930.

Ekbia, H., Mattioli, M., Kouper, I., Arave, G., Ghazinejad, A., Bowman, T., ... & Sugimoto, C. R. (2015). Big data, bigger dilemmas: A critical review. *Journal of the Association for Information Science and Technology*.

Etzkowitz H., & L. Leydesdorff. (2000). The dynamics of innovation: from National Systems and "Mode 2" to a Triple Helix of university–industry–government relations, *Research Policy*, *29*(2), 109-123.

Frické, M. (2014). Big data and its epistemology. *Journal of the Association for Information Science and Technology*, 66(4), 651-661.

Georgescu-Roegen, N. (1970). The economics of production. *The American Economic Review, 60*(2), 1-9.

Georgescu-Roegen, N. (1972). Process analysis and the neoclassical theory of production, *American Journal of Agricultural Economics,* 279-294.

Georgescu-Roegen, N. (1979). Methods in economic science, *Journal of Economic Issues,* 317-328.

Gruninger, M. & Fox, M.S. (1995). Methodology for the Design and Evaluation of Ontologies. In: *Proceedings of the Workshop on Basic Ontological Issues in Knowledge Sharing*, IJCAI-95, Montreal.

Huisman, J., Meek, V.L. & Wood, F.Q. (2007). Institutional diversity in higher education: a cross-national and longitudinal analysis, *Higher Education Quarterly,* 61/4: 563-577.

Imielinski T. & W. Lipski, Jr. (1984) Incomplete information in relational databases. *Journal of the ACM, 31*(4), 761–791.

Kitchin, R. (2014). Big Data, new epistemologies and paradigm shifts. *Big Data & Society, 1*(1), 1-12.

Lenzerini M. (2011). Ontology-based data management, *CIKM* 2011, 5-6.

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

Moed, W. Glanzel & U. Schmoch (ed.) (2004), *Handbook of Quantitative Science and Technology Research,* Kluwer Academic Publishers, 51-74.

Moed, H. F., & Halevi, G. (2015). The Multidimensional Assessment of Scholarly Research Impact, *Journal of the American Society for Information Science and Technology*, forthcoming.

Parnas, D. L. (1972). On the criteria to be used in decomposing systems into modules. *Communications of the ACM*, *15*(12), 1053-1058.

Poggi A. , D. Lembo, D. Calvanese, G. De Giacomo, M. Lenzerini, & R. Rosati (2008). Linking data to ontologies. *Journal on Data Semantics*, *10*,133–173.

REF (Research Excellence Framework) (2012). Panel Criteria and Working Methods. Retrieved January 7, 2015 from: http://www.ref.ac.uk/media/ref/content/pub/panelcriteriaandworkingmethods/01_12.pdf.

Sarma A. D., Dong X., & Alon Y (2008). Bootstrapping pay-as-you-go data integration systems. *Proc. of ACM SIGMOD*, 861–874.

Taylor, M. (2013). Exploring the Boundaries: How Altmetrics Can Expand Our Vision of Scholarly Communication and Social Impact. *Information Standards Quarterly*, *25*, 27-32.

Van Vught, F. (ed.) (2009). Mapping the higher education landscape: Towards a European classification of higher education. Dordrecht: Kluwer.

Van Vught, F., & Westerheijden, D.F. (2010). Multidimensional ranking: a new transparency tool for higher education and research, *Higher Education Management and Policy 22*/3, 1-26.

The Research Purpose, Methods and Results of the "Annual Report for International Citations of China's Academic Journals"

Junhong Wu, Hong Xiao[1,*], Shuhong Sheng[2,*], Yan Zhang, Xiukun Sun, Yichuan Zhang

[1, 2] *xh6613@cnki.net; SSH7600@cnki.net*

[1,2]Research Center of the Evaluation of the Scientific & Technical Literature of China, China National Knowledge Infrastructure (CNKI), 100192 Beijing (China)

Abstract

Before 2012, it was hard to come to a comprehensive evaluation of academic journals in China. For this reason the international influence of journals published in China hadn't been paid enough attention, leading to a bias in the Chinese research assessment system. Since 2012, China National Knowledge Infrastructure (CNKI) invested and carried out the project of the development of the "Annual Report for International Citations of Chinese Academic Journals ". In the same year, CNKI made a comprehensive study on the international citations of more than 6000 journals in China, and found that some journals had a certain international influence. In order to make a comprehensive assessment of the international influence of those journals, CNKI has developed a comprehensive indicator, named the CI index (clout index), combining the effect of both the impact factor and the citation counts. This article describes the purpose, methods and results of part of this project, providing a fresh idea for a comprehensive evaluation of the influence of Chinese academic journals.

Conference Topic

Methods and techniques

Background

In the era of big data and we-media as shown by Bowman & Willis (2003), direct publication and free access are all around, leading to the question: "how can academic journals survive"? It is known that journals, in particular journals sharing a scientific community compete in one market, but journals will survive as long as they have a function for a specific academic community. The main problem that Chinese journals, especially academic journals, are faced with, is the competition with huge international publishing companies. It has been a common knowledge that it is hard for the domestic journals to compete with those international academic journals.

According to Thomson Reuters' SCI data, as shown in Fig. 1, Chinese scholars published 114,130 papers in international journals in 2008. This number has greatly increased to 232.000 in 2013, which is a doubling of that in 2008. While Fig. 2 shows a comparison of the papers Chinese scholars published in the journals covered by the SCI with the papers Chinese scholars published in domestic academic journals in 2013. It can be seen in Fig. 2 that 1.035 million papers have been published in 3569 domestic academic journals in 2013. Compared to the 1,035,142 domestic papers, 206,598 academic papers were published in journals covered by the SCI. This means that one sixth of the Chinese academic papers had flowed overseas. There is also a rapid increase in quantity for the papers in the field of social sciences. The number of Chinese SSCI papers had increased from 4,430 in 2008 to 9,722 in 2013, which means a doubling over five years.

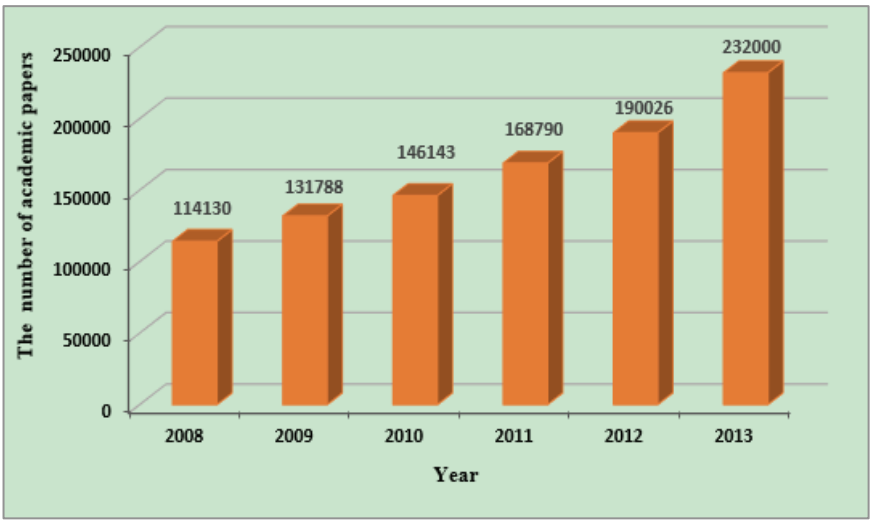


Figure 1. Evolution of the number of academic papers Chinese scholars published in international journals covered by the SCI during 2008-2013.

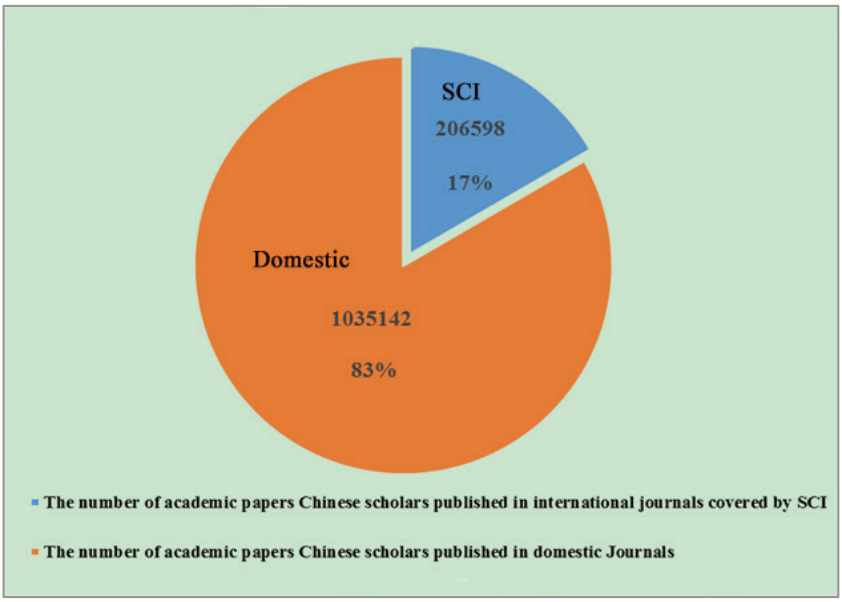


Figure 2. A comparison between Chinese papers published in journals covered by the SCI with papers published by Chinese scholars in domestic academic journals in 2013.

It is seen that so many China's qualified academic papers have flowed overseas and published in international journals, especially SCI journals. While the impact of China's academic journals on international audience was rarely revealed before. We think that, in the information era, with an increasing quantity of journals, the full importance of academic journals should be revealed through an objective evaluation based on a large amount of data. Journal management departments often use an "index set" of journal characteristics. This index set is used for the quantitative assessment of journal's quality. It has become a social consensus that "Scientific decision-making needs the support based on the big data". China's publishing management system, in particular, urgently needs a comprehensive, objective and impartial data set for the allocation of journal publishing resources.

Most scholars agree that publishing academic papers in the journals of a high academic standing is a means of academic communication and the success of a scholar in this can be used as a factor in evaluation exercises. The problem is where to draw the line between journals of high standing and journals of lower standing. In the past, data regarding

international journals were not taken into account for various evaluations of domestic journals. Therefore, it was hard for the domestic journals to compete with those international academic journals.

It is common that the reputation of Chinese scientific journals in the international community is measured by journal Impact factor (IF). Ren (1999) proposed the challenge for Chinese scientific journals using this indicator for the evaluation of Chinese journals. Based on the fact that a journal's international impact had not been adequately considered in the past, research management department such as that of CNKI could only take the SCI as evaluation standard. The Science Citation Index (SCI) initiated by Eugene Garfield is a unique retrieval and evaluation tool (Garfield, 1955). Yet it is known that it is not adequate for the local evaluations of less developed non-English speaking countries, or for the retrieval of these countries' publications (Ren & Rousseau, 2002). Since English is the most widely used language in science, journal publishers prefer to publishing in English to attract a larger reader base, resulting in more visibility, increased citations and higher IF, as shown in the study by Ren & Rousseau (2004). According to the above discussion, it is necessary for us to take an international perspective and a domestic view to evaluate the influence of China's academic journals, i.e. consider both domestic and international journals' citing citations to Chinese academic journals, in order to conduct scientific and reasonable evaluations of Chinese academic journals. For the citations by domestic journals, CNKI has developed "Annual Report for Chinese Academic Journal Impact Factors" since 2009. In this study, we focus on the citations by international journals, introducing the research purpose, methods and results of the "Annual Report for International Citations of Chinese Academic Journals".

Purpose

In this study, we conduct a quantitative assessment of academic journals published in Mainland China, either in Chinese or in English, in order to make an evaluation of their quality. Moreover, we analyze their world-wide influence by mining their cited records of citations by international journals. In the following we first give our understanding of an academic journal of high quality.

A journal of high quality provides products and services meeting or exceeding its readers' expectations. As such the quality of an academic journal is a comprehensive reflection of its publication level as manifested through the importance of its articles for the advancement of science. Following national and international norms, timeliness of publication and a large reader base also contribute to a journal's quality.

Influence of an academic journal refers to the ability of the journal to arouse its readers' attention and thinking, obtain their recognition and even alter their thoughts, opinions and behavior. A high-level journal influences academic development, by the ideas, concepts, theories, methods, findings, inventions and facts it introduces to the scientific community. Besides these objective aspects, a high-level journal has also an emotional influence, associated with its brand name, on its reader community.

Influence is not only a reflection of quality, but also a function of time. High quality papers, including editorials, show their influence gradually over time. Dissemination of journals can be judged scientifically and objectively by the frequency of being cited in domestic and foreign academic literature.

Method

Index system

Some scholars have suggested that data including downloading and online comments should be considered. These ideas are related to the altmetrics, or social influmetrics, movement

(Rousseau & Ye, 2013). Even the new Nature Index includes altmetric data (refer to the website: www.natureindex.com). However, downloading is a complex issue. It ranges from results of web crawling to students' learning, or providing intelligence services, and does not only include use for academic research. Moreover, based on current data analysis technology, it is still a challenge to judge if online comments are scientific or rigorous. In contrast, citation is a reflection of academic norms. Each author is required to respect the intellectual property rights of the literature he or she cites. Otherwise his/her behavior might be considered as misconduct. Therefore, statistical analysis of citations is considered as a relatively reliable and quantifiable technique.

Citation and publication statistics may include the following items:

(1) Statistics related to received citations such as the total cites in a year. Citations may include mutual journal citations and a history of received citations over a period of several years.

(2) Quantity of published literature such as the amount of published papers (further subdivided into types such as 'normal' articles, reviews, editorials, etc.), proportion of funded papers and proportion of articles with foreign collaboration.

(3) "Calculated indicators for evaluation: Indicators related with cited frequency such as immediacy index, the 2-year impact factor, 3-, 4-, or 5-year impact factor, etc. Indicator related with mutual citations: mutual citation index. Indicators related to the life cycle of literature: citing half-life, cited half-life, etc."

(4) The composition of the editorial board and the prestige of the editor-in-chief.

Selection of statistical sources for the international citations report

Statistical sources for the international citation report must be journals selected according to the standard for the evaluation of the international influence. Besides the journals from American and European countries, representative journals from other countries should also be included. The list of source journals should be based on suggestions from domestic and foreign experts. It is known that SCI database includes the most representative journals from the American and European countries and, as such, may be acceptable for reflecting the international influence of Chinese academic journals. Hence, at least for the current year, we still use the SCI database as the statistical source to evaluate academic journals. This means that we consider 8,621 academic journals covered in this database.

The case for humanities and social science fields is more complicated. It is not enough to merely use the 6,429 journals of SSCI and the A&HCI to evaluate the humanities and social sciences journals. For a more comprehensive statistic of the international influence of China's humanities and social sciences journals, we add well-known databases as a supplement, including those of leading international publishing groups such as Elsevier, Springer, Wiley, and Emerald. In this way, 1483 source journals (non-WOS humanities or social science journals) are included, which are good supplements for the source journals. According to experts' recommendations, we have also supplied 441 journals in minor languages, which pay attention to Chinese issues. These journals have not been included in the worldwide major databases, but they are indispensable for the research of local social science experts (Ossenblok, Engels & Sivertsen, 2012).

Data standards

In order to ensure the accuracy of the statistical data, we have established data processing standards, procedures, as well as quality requirements. Accordingly, we normalized and standardized the raw data and set up a series of databases as follows:

(1) The document database of norms for titles of more than 7,000 Chinese and English journals in China.
(2) The bibliographic database of China's academic journals, a collection of about 8,000 domestic academic journals and more than 42 million publications, used for citation links.
(3) Set up "the Statistical Standards of the Published Paper Amount". According the norm, make the statistics on the amount of published papers, as well as the cited papers published in the recent six years.

The development process

(1) Collection of data: including data retrieval in the WOS database, and the processing of data from supplementary journals.
(2) Standardized data processing: automatic processing of data such as citation links, and fuzzy title matching. If necessary these techniques were augmented by manual inspection to improve efficiency and accuracy.
(3) Detection: verification of data integrity and accuracy checks to ensure that the data meets the quality standards, plus an annual appraisal of a group of experts.
(4) Trial calculation, validation, and sample verification: indicator calculation must be double checked by several persons, and those journals with large inter-annual variations in one or more indices are the target of special attention.

Results

According to the method mentioned above, we developed the "International citation annual report" providing evaluation data of Chinese academic journals, first released in 2012. In December 2014, the 2014 Annual Report (Xiao & Du, 2014) was published and the evaluation data for more than 6000 academic journals were provided. These results were released in the "Database of Statistical Analysis of Individual Journal's Impact" available on the website of the CNKI (www.cnki.net).

Selection of highlights

Definition of a new international impact index: the clout index, denoted as CI

Several indicators like the journal impact factor and total cites are commonly used for the evaluation of a journal. Before continuing we first provide a short review of those indicators. The idea of a journal IF was first propagated by Eugene Garfield in the journal Science in 1955 (Garfield, 1955, 2006). Currently, the journal IF is generally regarded as representing the quality of academic journals in terms of citations received by its published articles. It is usually assumed that journals with a high IF carry meaningful, prominent, and quality research (Saxena, 2013). However, this single parameter is clearly not sufficient (Vanclay, 2012, Glaenzel, 2009). First of all, according to its definition, the IF reflects the performance of a journal in the most recent two years. The cited half-life for an academic journal is about 4 to 12 years, while the period of the most recent two years is merely the peak time for citation (and even that depends on the field), accounting for about 20% of the total citation amount, as shown in Fig.3. Second, the journal IF is independent of factors like the history and scale of journals and may reflect the popularity of published topics (Rousseau et al., 2013). Besides, journal IF depends on the research field: high journal IF is likely achieved for journals covering large areas of basic research with a rapidly expanding but short lived literature that use many references per article (Seglen, 1997). Using the IF as the single measure would lead to artificial constraints on the quantity of published work as well as the tendency to publish a large number of papers in line with the current popular trends without solving any

fundamental problem. Journals that act like this would lose their basic function as real academic communication platforms.

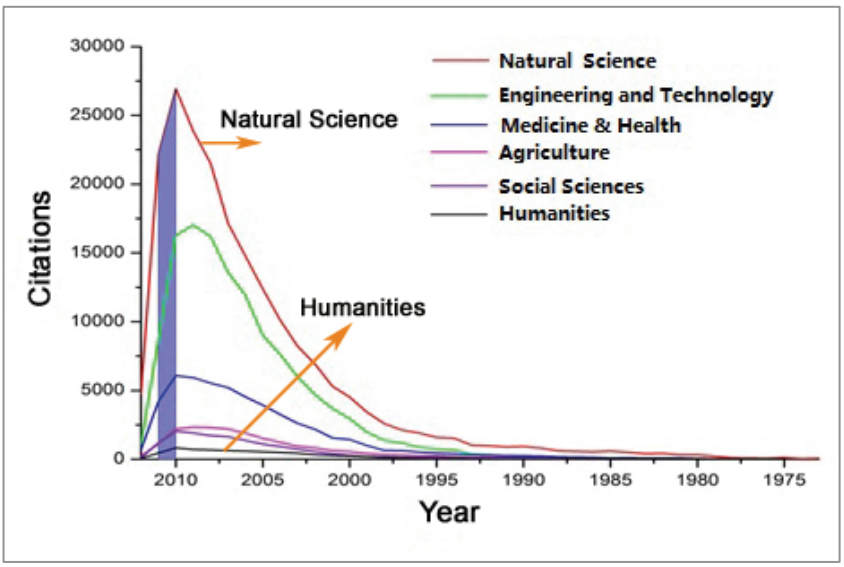


Figure 3. Cited half-life for different subject categories.

Total cites is directly related to journals' publishing history and scale. As shown in Fig. 3, we should also mention that total cites, which include the other 80% of citations that are excluded from the calculation of the IF, reflect the total impact power. However, it is also unreasonable to only consider total cites as the single evaluation factor. It might encourage researchers to aggressively increase their publishing quantity at the expense of academic quality.

Here, we should mention the topic of self citations. Whenever citations are used as indicators to evaluate scientific research, self-citations are often considered controversial. Many scholars have studied self-citation and some suggest that self-citations should be removed from citation counts, at least at micro and meso levels (e.g. analyses of persons, research groups, departments, and institutions) (Aksnes, 2003, Thijs & Glaenzel, 2005). Today the indicator of self-citations has been widely used in the evaluation of scientific journals.

We all know that IF and citations are field dependent, and therefore, indicators which compare expections to observed values are also interesting, see the work of Glaenzel, Schubert and Braun to MOCR (The Mean Observed Citation Rate) and MECR (The Mean Excepted Citation Rate) (Braun et al., 1985, Schubert et al., 1989). While Bonitz et al. (1997) stuied the Matthew effect of countries and Matthew citation journals, and presented the established characteristic of the so-called Matthew Effect for countries: field-dependency, time-stability and order of magnitude. Boyack & Klavans (2014) made the analysis on non-source publications in a different context, including non-source items in a large-scale map of science. These studies have inspired our work on exploring a comprehensive indicator for the evaluation of non-WOS-source domestic academic journals.

Thus, we have developed a comprehensive indicator, named as the Clout Index (CI), which takes both the IF and total cites into account. To be precise, we replace the WOS IF and total citations with the non-self-cited IF and total non-self-cited citations in the calculation of the CI values, taking into account that most of Chinese journals are not covered by the WOS.

First, we normalize the non-self-cited IF and total non-self-cited cites by a linear normalization method (the same for the two indices) shown in Equation (1), where V represents the parameter that has to be normalized, while N represents the normalized value. In this way the two values are in the range [0, 1].

$$N_i = \frac{V_i - V_{\min}}{V_{\max} - V_{\min}} \quad (1)$$

For the next step, we apply Eq. 2 to calculate the CI value:

$$CI = \sqrt{2} - C = \sqrt{2} - \sqrt{(1-x)^2 + (1-y)^2} \quad (2)$$

Fig. 4 shows a schematic distribution of CI values calculated by Eq.2. The points scattered in Fig. 4 represent the CI values of the selected journals. It should be noted that, in Eq. 2 and in Fig. 4, *x* and *y* stand for the normalized non-self-cited IF and total non-self-cited cites, respectively. From Fig.4, it can been seen that the origin coincides with non-self-cited IF = 0 and total non-self-cited cites = 0, while the point (1, 1) represents that the journal has reached the maximum value in both the non-self-cited IF and total non-self-cited cites. If we take the point (1, 1) as the center and CI value as radius to draw circles, then points on the same circle have the same CI value. The points located in the bottom-left area have lower CI values while the ones in the up-right area have higher values.

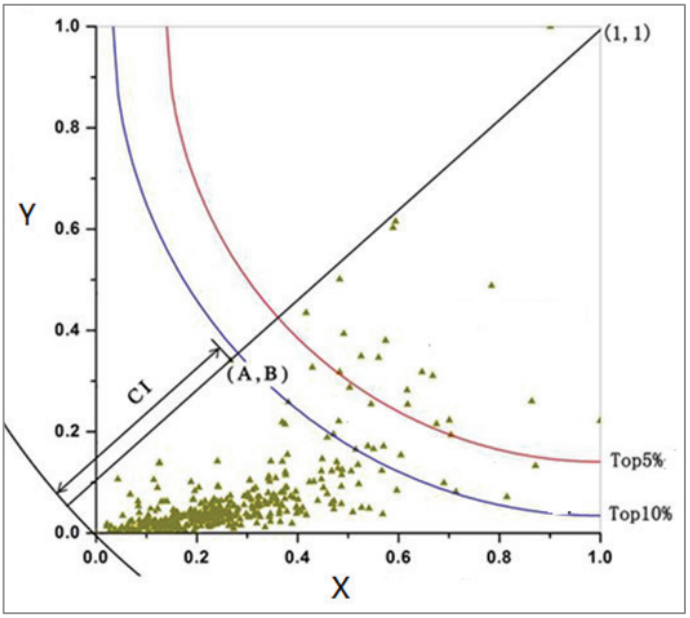


Figure 4. Schematic view of the Clout Index (CI) and selection of top journals in China.

In Figure 4, we drew two curves with the point (1,1) as their center, and CI(1) and CI(2) as radius, respectively. Here, CI(1) and CI(2), represent the critical CI values of selected Top 5% and Top 10% journals, respectively. We consider journals with CI values above the Top 5% as "The Highest International Impact Academic Journals of China", while CI values between CI(1) and CI(2) as "The Excellent International Impact Academic Journals of China". Here, China's academic journals published either in English or in Chinese are both considered.

At this point, we would like to explain why we choose a vector sum method for the calculation of the CI value instead of using a simple linear sum. Figure 5 shows a comparison of the linear sum and the vector sum method. The scatterplot itself in Figure 5 is the same as in Figure 4. First, we consider that IF and cites have the same weight as evaluation indicators. If we take the linear sum method, i.e. CI=*x*+*y*, we obtain oblique straight lines with different lengths to show equal CI values. Here, *x* and *y* have the same definition as those for Figure 5. Compared to the result obtained from the linear sum method, the CI value is smaller by the vector sum method, when the points closed to x or y axes. In this system, journals with higher single index, either IF or cites are easily excluded. Thus, our algorithm gives a better way to match the evaluation principle: "not only quality but also quantity matters".

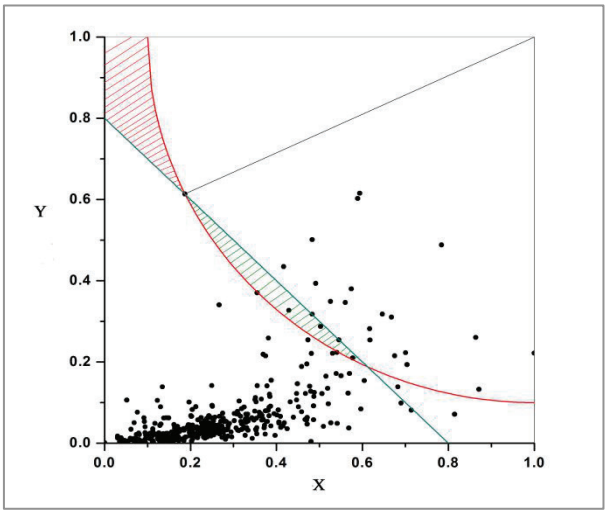


Figure 5. A comparison of the linear sum and the vector sum method.

In our system, we consider that journals with both higher IF and higher cites are journals of high-quality. Those journals commonly have higher influences in their scientific field. Fig. 6 shows the schematic view of the scatterplot of the CI value for both the SCI journals as well as Chinese domestic journals. We use double logarithmic coord. system in Fig. 6, where our developed vector sum method is applied to calculated the CI values for SCI journals and China's domestic journals. Both of the statistical sources are WOS database. The green triangles represent the CI values for the international SCI journals, while the black ones show the domestic journals covered by SCI. Those related to the Top 5% journals are shown in red, while for Top 5-10% journals in dark blue. Lower CI values for other domestic journal are shown in orange. It is clear that the majority of international journals covered by the SCI is situated in the area with both higher IF and higher citations. Situations are similar for the Top 5% and Top 5-10% domestic journals. Thus justifies that our vector sum method is a good method for the evaluation of the journals, and the CI index can be considered as an effective and reasonable indicator for the quantitative assessment of journal impact.

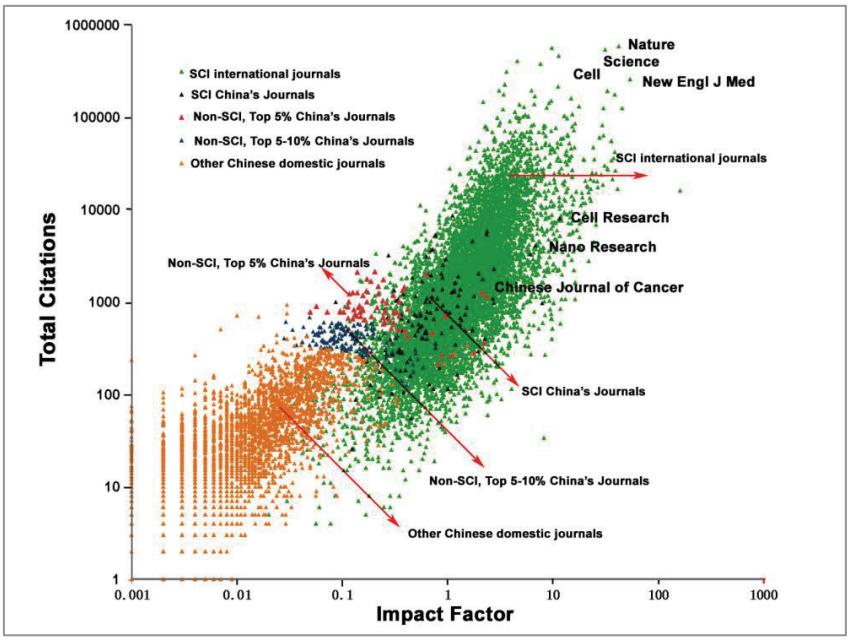


Figure 6. A schematic view of the scatterplot of the CI values for both the SCI journals and China's domestic journals.

Problems in the selection of top journals

Division of subjects

There is, among domestic journals in various disciplines, a large imbalance in the level of international influence and the extent of "going global". If top journals were selected per discipline, excellent journals in highly visible disciplines may be excluded, while journals with less international influence in disadvantaged disciplines may come into the list. This is not the way we want to highlight the most influential journals. As time goes by, we expect that the international influence of various disciplines will be improved and developed in a balanced way. At that time we will be able to perform sub-discipline rankings.

Comprehensive consideration of domestic and international influence

In the evaluation of international influence we should consider domestic and international influences as two sides of the same coin. However, first, a separate evaluation is more helpful for understanding the situation of the journals in both domestic and international market. Secondly, there is no recognized method showing how to merge these two reports, however, see Jin & Rousseau's study for an earlier merging attempts (Jin et al., 1999, Jin & Rousseau, 2004). The main difficulty lies in the point that there are different opinions on the issue of whether "a domestic citation is equal to an international citation." Considering the fact that "the annual report of the impact factors of China's academic journals" has been well developed for years, in this article, we mainly discuss the research method of the annual report of the international citations.

The selection process and resulting top list

To highlight the most influential journals, this year we continue selecting "The Highest International Impact Academic Journals of China" and "The Excellent International Impact Academic Journals of China". By ranking the journals of STM (Science/ Technology/ Medicine) and AH&SS (Humanities and Social Sciences) according to the CI values, we selected the Top 5% and the Top 5-10% journals; then sent the selection method, data of indicators and the primary list to more than 70 experts for peer review.

Some journals were removed from the list for their bad reputation evaluated by peer reviewers, while other ones were supplemented in sequence. This was done in such a way as to make sure that the total number of selected journals stayed the same. Finally we determined 176 STM journals and 61 AH&SS journals as "The Highest International Impact Academic Journals of China", and 174 STM journals and 60 AH&SS journals as "The Excellent International Impact Academic Journals of China" Among these 471 Top 5-10% journals, 458 are core journals selected by various domestic institutions, and most of the other 13 are journals in English or newly created ones.

Summary and outlook

(1) An international report has been issued for three successive years and approved by government departments in charge of journals, editorial department of journals and academic circles. This encourages us to keep this work going.

(2) With the accumulation of data, many meaningful conclusions can be drawn from the analysis of inter-annual variations in data.

(3) There is certainly room to improve the evaluation methods, including the selection criteria of the international source journals, possible improvement of the determination of the CI indicator and the integration of domestic and international lists.

(4) The main points of this article have been written in a manuscript submitted to Acta Editologica in Chinese language (Wu et al., 2014).

Acknowledgments

This study is sponsored by the National Social Science Project of "Evaluation and development strategy of the international influence of Chinese academic journals in English" (No.14BTQ055). The authors would like to express the deepest appreciation to Prof. Ronald Rousseau for his very careful corrections and valuable comments on this manuscript. We are very grateful to Prof. Jerold Mathews in Iowa State University for his careful corrections and good suggestions. And we thank Dr. Bo Liu in CNKI for his valuable discussion on the manuscript.

References

Aksnes, D.W. (2003). A macro study of self-citation. *Scientometrics*, *56*, 235–246.

Bonitz, M., Bruckner, E. & Scharnhorst, A. (1997). Characteristics and impact of the Matthew effect for countries, *Scientometrics 40*(3), 407-422.

Bowman, S. & Willis, C. (2003). We media-hypergene. Retrieved July 2003 from: http://www.hypergene.net/wemedia/download/we_media.pdf.

Boyack, K. W. & Klavans, R. (2014). Including cited non-source items in a large-scale map of science: What difference does it make? *Journal of Informetrics 8*(3), 569-580.

Braun, T., Glaenzel, W. & Schubert, A. (1985). *Scientometric indicators: A 32 country comparison of publication productivity and citation impact.* (pp. 424). Singapore: World Scientific Publishing Co.

Garfield, E. (1955). Citation indexes for science: a new dimension in documentation through association of ideas. *Science*, *122*, 108-111.

Garfield, E. (2006). The history and meaning of the journal impact factor. *JAMA*, *295*, 90-93.

Glaenzel, W. (2009). The multi-dimensionality of journal impact. *Scientometrics*, *78*(2), 355-374.

Jin, B., Wang, S., Wang, B. et al. (1999). A unified method of counting international and domestic articles. J*ournal of Management Sciences in China*, *2*(3), 59–65.

Jin, B.H. & Rousseau, R. (2004). Evaluation of Research Performance and Scientometric Indicators in China. In H. F. Moed, W. Glänzel & U. Schmoch (Eds.), *Handbook of Quantitative Science and Technology Research* (pp. 497-514). Dordrecht, etc.: Kluwer Academic Publishers.

Ossenblok, Truyken L.B., Engels, Tim C. E. & Sivertsen, G. (2012). The representation of the social science and humanities in the Web of Science: a comparison of publication patterns and incentive structures in Flanders and Norway (2005-9). *Research Evaluation*, 21(4), 280-290.

Ren, S.L., Liang, P. & Zu, G. (1999). The challenge for Chinese scientific journals. *Science*, *286*, 1683.

Ren, S.L. & Rousseau, R. (2002). International Visibility of Chinese scientific journals. *Scientometrics*, *53*, 389-405.

Ren, S.L. & Rousseau, R. (2004). The role of China's English-language scientific journals in scientific communication. *Learned Publishing*, *17*, 99-104.

Rousseau, R., Garcia-Zorita, C. & Sanz-Casado, E. (2013). The h-bubble. *Journal of Informetrics*, *7*(2), 294-300.

Rousseau, R. & Ye, Fred Y. (2013). A multi-metric approach for research evaluation. *Chinese Science Bulletin*, *58*, 3288-3290.

Saxena, A., Thawani, V., Chakrabarty, M. & Gharpure, K. (2013). Scientific evaluation of the scholarly publications. *Journal of Pharmacology and Pharmacotherapeutics*, *4*(2), 125-129.

Schubert, A., Glaenzel, W. & Braun, T. (1989). World flash on basic research: scientometric datafiles. A comprehensive set of indicators on 2649 journals and 96 countries in all major science fields and subfields, 1981–1985. *Scientometrics*, *16*(1–6), 3–478.

Seglen, Per O. (1997). Why the impact factor of journals should not be used for evaluating research. *BMJ*, 314, 497-502.

Thijs, B. & Glaenzel, W. (2005). The influence of author self-citations on bibliometric meso-indicators. The case of European universities. *Scientometrics*, *66*, 71–80.

Vanclay, J.K. (2012). Impact factor: outdated artefact or stepping-stone to journal certification? *Scientometrics*, *92*(2), 211-238.

Wu, J.H., Xiao, H., Zhang, Y., et al. (2014). A comprehensive index algorithm of the international influence evaluation of Chinese academic journals, submitted to *Acta Editologica*.

Xiao, H. & Du, W. T. (Eds.) (2014). Annual Report for International Citation of Chinese Academic Journals. Beijing: "China Academic Journals (CD-ROM Version)" Electronic Publishing House.

Is the Year of First Publication a Good Proxy of Scholars' Academic Age?

Rodrigo Costas[1], Tina Nane[2] and Vincent Larivière[3]

[1] *rcostas@cwts.leidenuniv.nl;* [2] *g.f.nane@cwts.leidenuniv.nl*
Centre for Science and Technology Studies (CWTS), Leiden University, P.O. Box 905, 2300 AX, Leiden (the Netherlands)

[3] *vincent.lariviere@umontreal.ca*
École de bibliothéconomie et des sciences de l'information, Université de Montréal, C.P. 6128, Station Centre-Ville Montreal, Quebec (Canada)

Abstract
Individual scholars are the central unit of the research system and are increasingly the focus of bibliometric studies. An important aspect in the study of individual scholars is their academic age, which allows for the comparison of scholars that have been academically active in a similar period of time. Based on a sample of Quebec researchers for whom their year of birth, PhD year as well as the year of their first publication are known, we study the relationships among these ages with the aim of determining how their year of first publication can be used to estimate their 'real' age. Moderate correlations have been found among the ages, and the first publication year has a higher correlation with the PhD year than with the birth year. However, an important dispersion of scholars across the different ages is observed; thus, the year of first publication can only be taken as proxy of the real age of scholars. Alternatively, the consideration of cohorts of around 5 years seems to be a reasonable approach. Further research will focus on the exploration of other bibliometric indicators in order to refine the preliminary developments discussed here.

Conference Topic
Methods and techniques

Introduction
In individual-level bibliometric studies, the socio-demographic characteristics of scholars are of central importance to understand and better frame the results obtained (Costas & Bordons, 2011; Gingras, Larivière, Macaluso, & Robitaille, 2008; Mauleón & Bordons, 2006). Among these socio-demographic characteristics we can mention gender (Larivière, Ni, Gingras, Cronin, & Sugimoto, 2013; Mauleón & Bordons, 2006), mobility (Canibano, Otamendy, & Solis, 2011; Franzoni, Scellato, & Stephan, 2012), and nationality (Moed & Halevi, 2014), among others. The development of large-scale author-name disambiguation algorithms (Caron & Van Eck, 2014) as well as the increasing quantity of papers' metadata indexed (e.g. author names and surnames, affiliations, e-mail data, etc.) have allowed the study of the socio-demographic characteristics of scholars at a larger scale. For example, the analysis of the first author names of authors (Larivière et al., 2013) allowed the macro analysis of gender disparities worldwide. The large-scale analysis of the relationship between author names, affiliations and countries collected from scientific publications has open the possibility of studying academic mobility at the world level (Moed, Aisati, & Plume, 2013), as well as the nationality (Costas & Noyons, 2013), migrations (Moed & Halevi, 2014) or even the ethnic origin (Freeman, 2014) of scholars.

A critical element for individual-level bibliometrics is the age of the researchers (Costas & Bordons, 2011; Larivière, Archambault, & Gingras, 2008; Levin & Stephan, 1989), especially from the point of view of its relationship with productivity (Falagas, Ierodiakonou, & Alexiou, 2008; Levin & Stephan, 1989). Age is also a common point of debate in science policy, as it aims to compare scholars of the same 'academic age' (Bornmann & Leydesdorff,

2014). However, one of the main reasons that hinders the development of bibliometric studies at the individual level is the lack of systematic data on the age of scholars, as this information is not systematically collected in bibliographic databases. A commonly used proxy for the study of the age of scholars has been the so-called 'scientific (or academic) age', often defined as the publication year of the first paper of a scholar (Radicchi & Castellano, 2013).[1] The use of this age is very convenient, as it is possible to directly extract it from bibliometric data. However, so far there has not been any analysis on the relationship between this proxy and the real age of scholars. This paper is intended to fill this gap and shed some light on the relationship between the 'bibliometric' age of scholars that can be calculated based on bibliographic information and the 'real' age(s) of individual scholars, namely their birth age and their PhD age. In other words, we aim to infer the birth year and PhD year of scholars based on models that are exclusively based on bibliometric indicators[2] (e.g. first publication year, position of signature, co-authors, etc.). Thus, the main research question can be operationalized as follows: *could the year of first publication (YFP) of a scholar (as recorded in the Web of Science) be considered as a relevant proxy of the birth and/or PhD ages of scholars?*

Methodology

In order to answer the research questions it is necessary to have a dataset of scholars for whom their real ages are certainly known as well as the publication years of their scientific publications. Thus, as our golden set, in this study we have considered one of the (possibly) largest datasets of individual scholars for whom real individual characteristics are known (this dataset has been used in some other studies, e.g. Gingras et al., 2008; Larivière et al., 2011). This dataset is composed by 13,626 university professors from Quebec who have published at least one article during the 1980-2012 period. For every scholar in the dataset, the following individual elements have been codified:

- Year of birth [*Birth year*]
- Year of PhD (year when the scholar has obtained her (first) PhD) [*PhD year*].
- Publication year of their first publication in the Web of Science (WoS) [*YFP*]
- [*Birth year to YFP*], which is calculated as [*YFP*]-[*Birth year*]
- [*PhD year to YFP*] which is calculated as [*YFP*]-[*PhD year*]
- Domain (nine disciplinary fields of activity of the scholar, which is based on the 2000 revision of the U.S. Classification of Instructional Programs (CIP)[3] developed by the U.S. Department of Education's National Center for Education Statistics (NCES).

Complementary, we have also calculated the total number of publications of the scholars in the period 1980-2012 [*P*].

A technical limitation of the dataset is that the WoS publication data starts in 1980, thus meaning that for very old individuals it is not possible to know with certainty if the first publication recorded in the WoS during the period 1980-2012 truly corresponds with their actual first publication. To reduce the effect of this issue, we decided to focus only on those individuals that have a birth year later than 1959 (i.e. we don't expect that many scholars would have a publication before their 20's) and a PhD year also later than 1980 (same criteria

[1] Although this term has also been proposed for the time since the PhD has been awarded (Bar-Ilan, 2014). Some other studies have also focused on the starting year of publication of individuals as proxies of age (Fronczak, Fronczak & Holyst, 2006).

[2] Due to space restrictions, in this paper we focus only on the first publication year as a proxy, and leave for a further version of this paper the consideration of other bibliometric variables.

[3] The Classification of Instructional Programs (CIP) is developed by the U.S. Department of Education's National Center for Education Statistics (NCES). More details can be found at: http://nces.ed.gov/pubs2002/cip2000/

as before). As a result of this filtering we ended up with 3,596 scholars that are the final dataset of our analysis.

Main results

This section presents the main results of the analysis. In Appendix 1 the descriptive scores are presented. Results show that there are differences in individual productivity by domain, which is of course not a surprise. For instance scholars from the Basic Medical Sciences and Health sciences exhibit the highest number of WoS papers, while Humanities the lowest. Similarly, the median birth year of the whole sample is 1965, although there are small differences by domain, with Basic Medical Sciences with the oldest individuals (median=1964) and Social Sciences the youngest (median=1967). The median PhD year of the whole sample is 1998, with the Basic Medical Sciences as the oldest median (1994) and domains such as Business & Management, Education, Non-health professionals getting their PhD on median in 1998.

Regarding the time between the birth of the scholars and the time of their first publication, scholars from Basic Medical Sciences, Engineering, Health Sciences and Science are on median the fastest (32 years) while scholars from Business & Management, Education or Humanities are slower (35 years). From the PhD to the first publication, the fastest are the scholars in Health Sciences (1 year) and the slowest the Humanities (4 years). It is important to keep in mind that here we also have cases with negative values, which means that researchers publish publications before their PhD date; a finding coherent with Larivière (2012).

Relationship between the different ages

In Appendix 2 we present the main correlations between the different ages of the scholars. In Figure 1 a summary of the correlations is presented. In general, there is a reasonably good correlation between birth year and PhD year, and the two real ages of the scholars have moderate correlations with YFP, although the PhD year has a generally better correlation with YFP than the Birth Year. These results suggest that it is reasonable to consider the YFP as a proxy of the scientific age of the researchers.

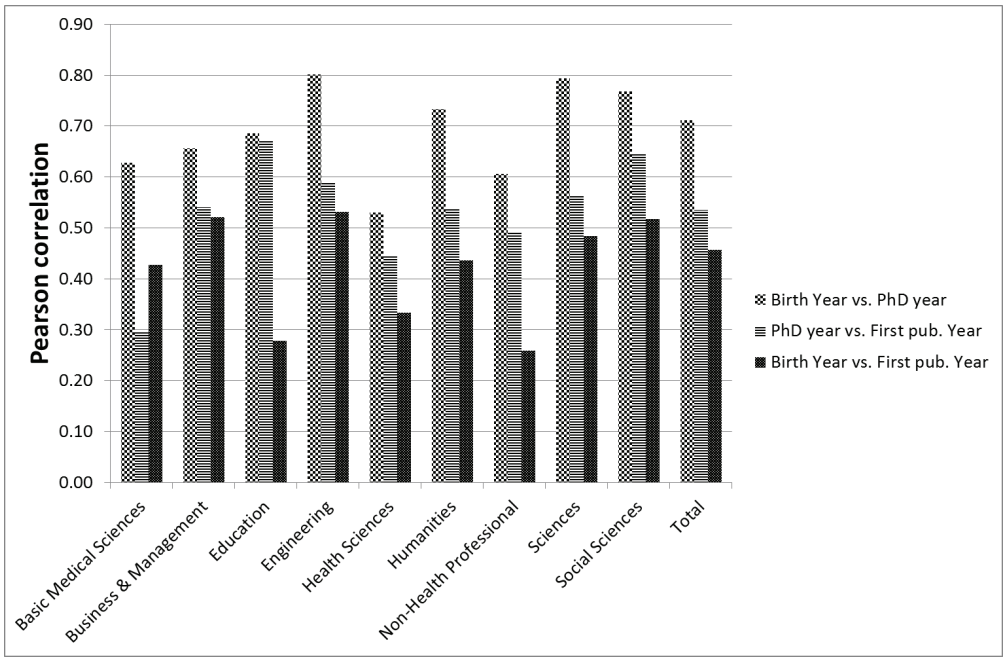


Figure 1. Pearson correlation values of the different ages – by disciplines and all disciplines combined.

YFP as a proxy of the age of researchers

Considering the moderate correlations between the YFP and the real ages of the researchers, we explore the dispersion of the scholars by the different ages. In Figure 2 box plots of each of the three variables (YFP, Birth year and PhD year) grouped by the combination of the same variables are presented. Thus it is possible to understand how scholars distribute across the different ages.

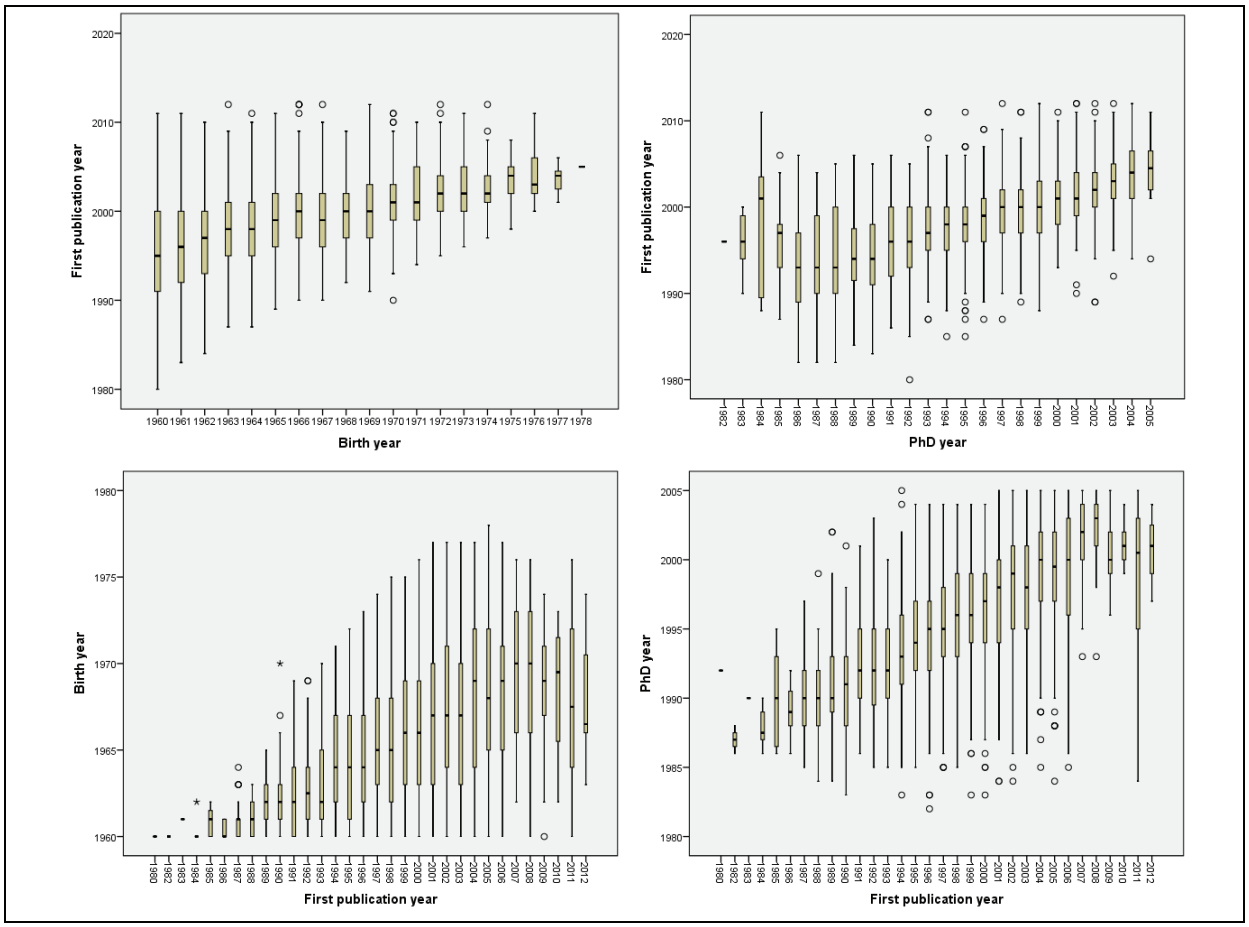


Figure 2. Box plot distribution of scholars across the different ages (all scholars together).

The two graphs on top of Figure 2 present boxplots of YFP observations grouped by each distinct birth year and PhD year. In the case of the birth year, it is possible to see how the earlier the year of birth the larger the variation of the YFP, thus indicating how researchers of all ages start their publication activities at different points in their lives, although the majority (i.e. the 'box' in the graph) tends to concentrate in a range of 5 to 10 years. The YFP median also tends to increase as the birth year increases. In the case of the PhD year we see also a quite disperse distribution of the first publication year of the scholars, although (with the exception of some irregularities among the scholars with the earliest PhD years) we notice a stepper increase in the median value of the YFP as the PhD year increases.

The graphs on the bottom of Figure 2 show the distribution of the two real ages (birth and PhD years) as a function of the YFP. Here we can also see an important dispersion of scholars across the two ages. However, in order to summarize the results of these two graphs, in Figure 3 we present the interquartile ranges (i.e. range of the number of years that include the 50% of all the observations), thus allowing to identify where most of the scholars are located in the distribution as a function of their first publication year.

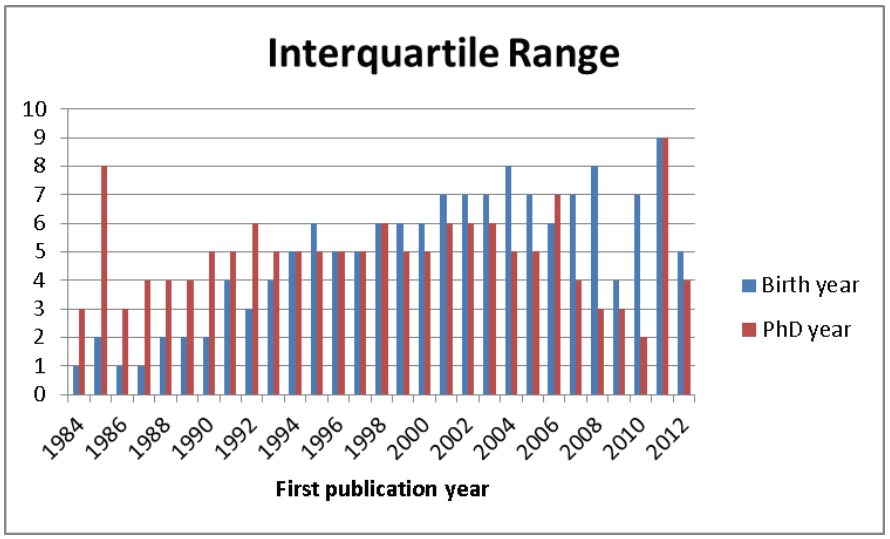


Figure 3. Interquartile range (in number of years) for Birth year and PhD year as a function of YFP.

Figure 3 shows that the interquartile range in all cases is smaller than 10 years for any of the two ages considered. Actually the average for all the YFP years considered is 4.9 years for both ages (with a median of 5). Thus, a possible interpretation of this result is that if we would only count with the YFP of the scholars, with a range of around 5-10 years we would be able to capture the real age of about 50% of all the scholars who started to publish that year.

Exploring a predictive model for the age of scholars based on bibliometric indicators

In this section a more predictive approach is presented. We are interested in estimating the birth and PhD years of a generic researcher by using the YFP indicator in our data sample. Numerous approaches can be taken, from the selection of different models and independent variables that could influence the two ages. In the present study we choose the simple linear regression model, with the average birth year and the average PhD year as dependent variables and the YFP as the independent variable. We will therefore infer on the average birth and PhD year of a scholar, and Figures 4 and 5 provide the linear regression fit of the two models, along with confidence and prediction intervals.

Using linear regression analysis the average ages (birth year and PhD year) of the whole list of scholars are fitted, including a 95% confidence interval as well as a 95% prediction interval. Although both intervals account for the uncertainty of the regression parameter estimates, there is an important distinction between the two intervals. The confidence interval is supposed to cover the true average birth year (of all the scholars in the statistical population) with high probability in 95% of the cases. The prediction interval provides limits on a future sampled observation that is an average of a given number of scholars from the set of all the scholars in the world. The prediction intervals refer then to actual observations in the data, and hence account also for the variation in the data, whereas the confidence intervals refer to the population's (of all scholars) average birth year. The prediction intervals are always larger than the confidence intervals.

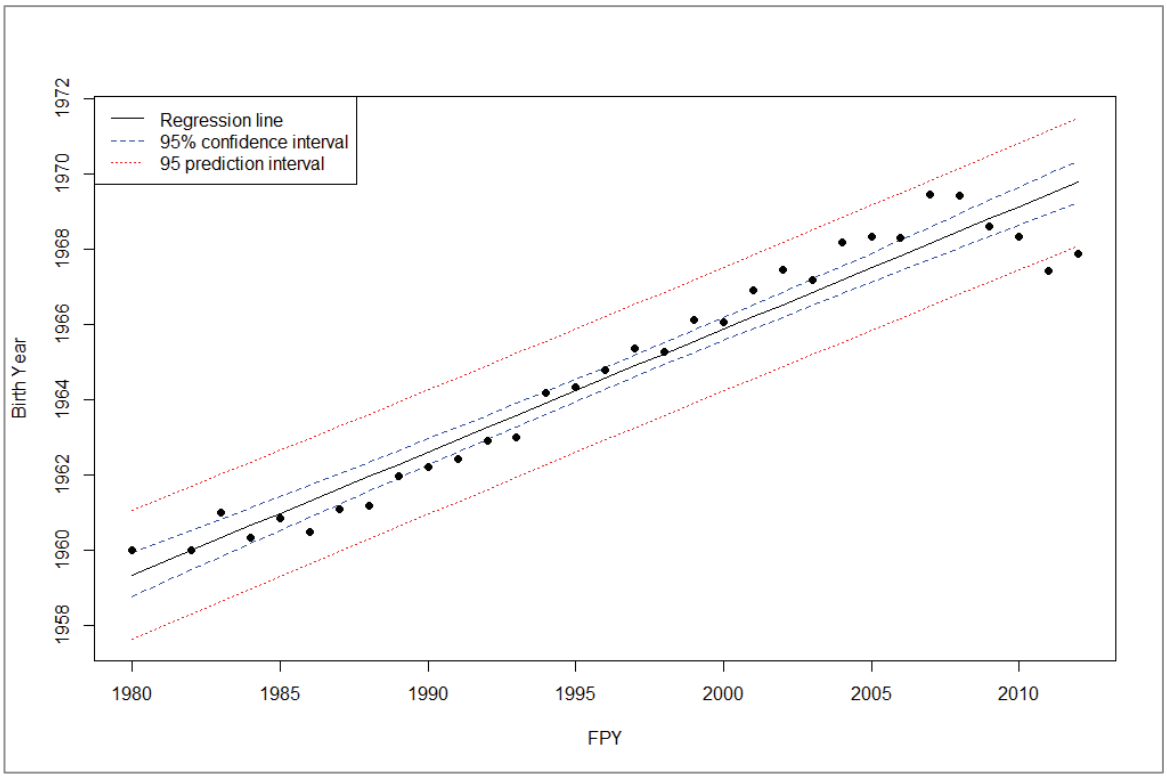


Figure 4. Average birth year by YFP, fitting a regression line and 95% confidence and prediction intervals.

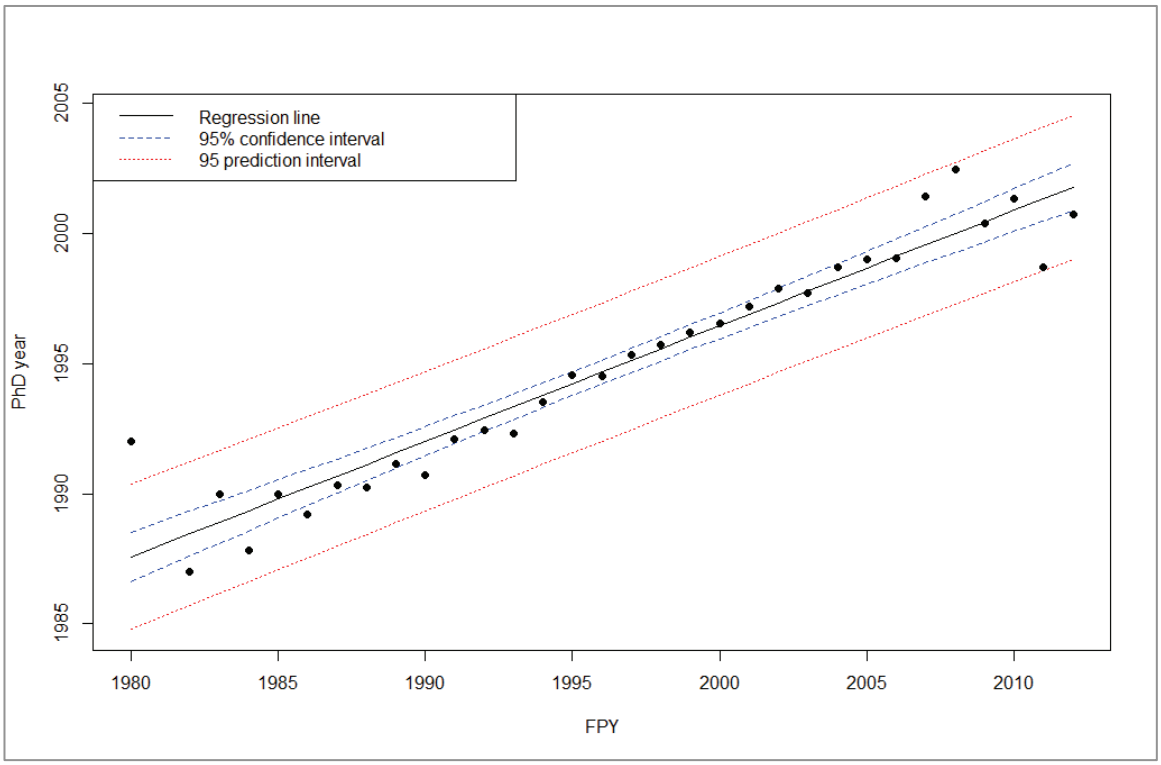


Figure 5. Average PhD year by YFP, fitting a regression line and 95% confidence and prediction intervals.

The main difference with bottom graphs in Figure 2 is that here the target is to estimate the average age of scholars from a given YFP. For example, in Figure 4 we can see how for scholars with a YFP=1995 their average birth year would be 1963, and the prediction interval ranges between 1961 and 1965. A similar pattern is observed in Figure 5 for PhD year (i.e.

with YFP=1995 the average PhD year would range around a period of five years). This suggests that we would be able to estimate the average ages of the scholars with a given YFP within an interval of 5 years. Of course, it is important to keep in mind that this analysis is based on the average values for all scholars, which is different from the individual prediction of individual scholars; however the relatively short prediction intervals (around 5 years) supports the importance of the YFP as relevant proxy for the ages of individual scholars.

Discussion and conclusions

Age is one of the most important socio-demographic determinants of researchers' activities, funding, output and impact. However, the lack of systematically recorded information on the age (real or academic) of researchers makes the need of developing reliable and valid proxies a priority. So far, the age of the first publication of individual scholars has been frequently considered as a proxy of the real age of scholars; however its validity has never been tested. Based on a sample of Quebec researchers for whom their actual birth year, PhD year as well as the year of their first publication are known, a study on the relationships among these ages has been performed.

The three ages correlate moderately well, birth year and PhD year have a good relationship, and YFP has moderate correlations with the other two ages, particularly with the PhD year. It is also possible to detect an important dispersion of scholars across the different ages, indicating that new authors (and new researchers) basically can come from a wide range of years. This means that, in spite of the moderate correlation between the YFP and the other ages, the YFP can only be considered as a proxy for researchers' age, as it does mix researchers with different birth and PhD years. The consideration of cohorts of years seems to be a more reasonable alternative. Thus, it is possible to argue that considering authors who started to publish in a given year, the majority of these scholars would have ages (birth and PhD) within a range of 5 to 10 years.

It is important also to highlight some of the limitations of this study. In the first place, we are working with a dataset of scholars from only one location (Quebec in Canada), so we need to keep in mind the limitations of the representativeness of our sample for the whole world. Thus, issues related with the changes and internal evolution of PhD programs could partly influence the results and hinder their generalization. Secondly, WoS is the only database considered for the determination of the YFP, however scholars can publish outputs not covered by this database, which is likely the case in Quebec, whose local literature in the social sciences and humanities is highly relevant (Larivière & Macaluso, 2011). Thirdly, in this study we haven't explored differences across fields, but arguably there are differences in the relationship between the ages and the first publication year of the scholars as disciplinary differences in individual productivity have been also discussed (Ruiz-Castillo & Costas, 2014).

All in all, considering the limitations previously exposed, our results are still policy-relevant and support the idea that the first publication year(s) of individual scholars can work as a reasonable proxy as their age, particularly when considering cohorts of researchers. For the final version of the paper other approaches will be also considered, including the analysis of the positions of the scholars in the papers (as these positions are related with the age of scholars (Costas & Bordons, 2011), other bibliometric indicators (e.g. the total number of publications of a scholar and total number of citations, which are age dependent) as well as the different disciplines of scholars. Finally, the consideration of other datasets from other countries and/or disciplines is an important development in order to globally validate the different tests and models obtained and to establish a more generalizable approach for the estimation of ages based on bibliometric data. A potential recommendation derived from this study is the relevance of incorporating information about the age, PhD year, gender and other

demographic characteristics in modern Research Information Systems (RIS). This would allow for more accurate studies of the demographics and changes in the trends of scientific productivity of individual scholars.

References

Bar-Ilan, J. (2014). Evaluating the individual researcher - adding an altmetric perspective. *Research Trends*, *37*, 31–34.

Bornmann, L. & Leydesdorff, L. (2014). On the meaningful and non-meaningful use of reference sets in bibliometrics. *Journal of Informetrics*, *8*(1), 273–275. doi:10.1016/j.joi.2013.12.006

Canibano, C., Otamendy, F. J. & Solis, F. (2011). International temporary mobility of researchers: a cross-discipline study. *Scientometrics*, *89*(2), 653–675.

Caron, E. & Van Eck, N. J. (2014). Large scale author name disambiguation using rule-based scoring and clustering. In E. Noyons (Ed.), *19th International Conference on Science and Technology Indicators. "Context counts: pathways to master big data and little data."* Leiden: CWTS-Leiden University.

Costas, R. & Bordons, M. (2011). Do age and professional rank influence the order of authorship in scientific publications? Some evidence from a micro-level perspective. *Scientometrics*, *88*(1), 145–161. Retrieved from http://www.springerlink.com/index/10.1007/s11192-011-0368-z

Costas, R. & Noyons, E. (2013). Detection of different types of " talented " researchers in the Life Sciences through bibliometric indicators : methodological outline Sciences through bibliometric indicators : methodological outline 1. *CWTS Working Paper Series*, (CWTS-WP-2013-006). Retrieved from http://www.cwts.nl/pdf/CWTS-WP-2013-006.pdf

Falagas, M. E., Ierodiakonou, V. & Alexiou, V. G. (2008). At what age do biomedical scientists do their best work? *The FASEB Journal*, *22*(12), 4067–4070.

Franzoni, C., Scellato, G. & Stephan, P. (2012). Patterns of international mobility of researchers : evidence from the GlobSci survey. In *International Schumpeter Society Conference* (pp. 1–32). Retrieved from http://www.aomevents.com/media/files/ISS 2012/ISS SESSION 7/Scellato.pdf

Freeman, R. B. (2014). Strength in diversity. *Nature*, *513*, 305.

Fronczak, P., Fronczak, A. & Holyst, J. A. (2006). Publish or perish: analysis of scientific productivity using maximum entropy principle and fluctuation-dissipation theorem. *arXiv*.

Gingras, Y., Larivière, V., Macaluso, B. B., Robitaille, J.-P. & Lariviere, V. (2008). The Effects of aging on researchers' publication and citation patterns. *Plos ONE*, *3*(12), e4048. doi:10.1371/journal.pone.0004048

Lariviere, V., Archambault, E. & Gingras, Y. (2008). Long-term variations in the aging of scientific literature: from exponential growth to steady-state science (1900-2004). *Journal of the American Society for Information Science and Technology*, *59*(2), 288–296. doi:10.1002/asi

Larivière, V., Gingras, Y., Cronin, B. & Sugimoto, C. R. (2013). Global gender disparities in science. *Nature*, *504*, 4–6.

Levin, S. G. & Stephan, P. E. (1989). Age and research productivity of academic scientists. *Research in Higher Education*, *30*(5), 531–549.

Mauleón, E. & Bordons, M. (2006). Productivity , impact and publication habits by gender. *Scientometrics*, *66*(1), 199–218.

Moed, H. F., Aisati, M. M. & Plume, A. (2013). Studying scientific migration in Scopus. *Scientometrics*, *94*, 929–942. doi:10.1007/s11192-012-0783-9

Moed, H. F. & Halevi, G. (2014). A bibliometric approach to tracking international scientific migration. *Scientometrics*, 1987–2001. doi:10.1007/s11192-014-1307-6

Radicchi, F. & Castellano, C. (2013). Analysis of bibliometric indicators for individual scholars in a large data set. *Scientometrics*, *97*(3), 627–637. doi:10.1007/s11192-013-1027-3

Ruiz-Castillo, J. & Costas, R. (2014). The skewness of scientific productivity. *Journal of Informetrics*, *8*(4), 917–934. doi:10.1016/j.joi.2014.09.006

Appendix 1. Main descriptive values

Disciplinary division		Birth year	PhD year	YFP	P	Birth year to YFP	PhD year to YFP
Basic Medical Sciences	N	713	713	713	713	713	713
	Mean	1993.66	1964.72	1997.01	52.54	32.29	3.34
	Std. Deviation	4.503	3.427	4.835	67.27	4.58	5.54
	Median	1994.00	1964.00	1997.00	30.00	32	3
	Minimum	1983	1960	1980	1	20	-13
	Maximum	2005	1976	2008	788	46	21
Business & Management	N	243	243	243	243	243	243
	Mean	1997.50	1965.92	2000.56	10.92	34.64	3.05
	Std. Deviation	4.427	4.313	4.476	12.405	4.31	4.269
	Median	1998.00	1965.00	2001.00	7.00	35	3
	Minimum	1983	1960	1986	1	25	-10
	Maximum	2005	1976	2012	96	49	27
Education	N	47	47	47	47	47	47
	Mean	1997.38	1965.49	2001.04	8.43	35.55	3.66
	Std. Deviation	3.943	4.117	5.254	13.333	5.71	3.93
	Median	1998.00	1965.00	2001.00	4.00	35	3
	Minimum	1989	1960	1986	1	25	-5
	Maximum	2003	1974	2010	70	48	12
Engineering	N	514	514	514	514	514	514
	Mean	1996.38	1966.27	1998.67	38.08	32.40	2.30
	Std. Deviation	4.713	4.488	4.509	48.889	4.36	4.19
	Median	1996.00	1966.00	2000.00	24.50	32	2
	Minimum	1982	1960	1985	1	22	-11
	Maximum	2005	1977	2009	692	44	17
Health Sciences	N	292	292	292	292	292	292
	Mean	1996.89	1965.45	1998.10	49.80	32.65	1.20
	Std. Deviation	4.183	4.006	4.800	72.488	5.13	4.76
	Median	1997	1965	1998	30	32	1
	Minimum	1985	1960	1984	1	22	-13
	Maximum	2005	1976	2012	788	49	18
Humanities	N	347	347	347	347	347	347
	Mean	1996.78	1965.76	2001.11	3.91	35.35	4.32
	Std. Deviation	4.341	4.115	4.382	5.338	4.52	4.19
	Median	1997	1965	2001	2	35	4
	Minimum	1986	1960	1986	1	24	-6
	Maximum	2005	1978	2012	65	47	20
Non-Health Professional	N	112	112	112	112	112	112
	Mean	1997.84	1965.52	2001.21	10.30	35.70	3.36
	Std. Deviation	4.594	4.480	5.070	14.222	5.84	4.89
	Median	1998	1965	2001.5	4	35	3

Disciplinary division		Birth year	PhD year	YFP	P	Birth year to YFP	PhD year to YFP
	Minimum	1985	1960	1990	1	24	-6
	Maximum	2005	1977	2012	70	51	21
Sciences	N	826	826	826	826	826	826
	Mean	1995.35	1965.88	1997.92	36.45	32.04	2.57
	Std. Deviation	4.441	4.287	4.860	48.406	4.67	4.37
	Median	1996	1965	1999	25.00	32	3
	Minimum	1985	1960	1982	1	22	-11
	Maximum	2005	1977	2012	775	46	17
Social Sciences	N	502	502	502	502	502	502
	Mean	1997.36	1966.75	1999.66	15.87	32.9084	2.3008
	Std. Deviation	4.25	4.33	4.53	19.11	4.36	3.7
	Median	1998.00	1967.00	2000.00	10.00	33	2
	Minimum	1987	1960	1986	1	23	-11
	Maximum	2005	1977	2012	204	48	15
Total	N	3596	3596	3596	3596	3596	3596
	Mean	1995.95	1965.77	1998.73	32.04	32.97	2.78
	Std. Deviation	4.64	4.18	4.89	50.56	4.77	4.60
	Median	1996	1965	1999	17	33	3
	Minimum	1982	1960	1980	1	20.00	-13.00
	Maximum	2005	1978	2012	788	51.00	27.00

Appendix 2. Pearson correlations by ages

Division	Ages	Birth year	YFP	PhD year
Basic Medical Sciences	Birth year	1.000	0.426	0.627
	First publication year	0.426	1.000	0.297
	PhD year	0.627	0.297	1.000
Business & Management	Birth year	1.000	0.521	0.656
	First publication year	0.521	1.000	0.540
	PhD year	0.656	0.540	1.000
Education	Birth year	1.000	0.277	0.686
	First publication year	0.277	1.000	0.670
	PhD year	0.686	0.670	1.000
Engineering	Birth year	1.000	0.531	0.800
	First publication year	0.531	1.000	0.588
	PhD year	0.800	0.588	1.000
Health Sciences	Birth year	1.000	0.333	0.530
	First publication year	0.333	1.000	0.444
	PhD year	0.530	0.444	1.000
Humanities	Birth year	1.000	0.435	0.733
	First publication year	0.435	1.000	0.538
	PhD year	0.733	0.538	1.000
Non-Health Professional	Birth year	1.000	0.258	0.605
	First publication year	0.258	1.000	0.492
	PhD year	0.605	0.492	1.000
Sciences	Birth year	1.000	0.484	0.793
	First publication year	0.484	1.000	0.561
	PhD year	0.793	0.561	1.000
Social Sciences	Birth year	1.000	0.517	0.768
	First publication year	0.517	1.000	0.646
	PhD year	0.768	0.646	1.000
Total	Birth year	1.000	0.457	0.711
	First publication year	0.457	1.000	0.535
	PhD year	0.711	0.535	1.000

Corpus Specific Stop Words to Improve the Textual Analysis in Scientometrics

Vicenç Parisi Baradad[1] and Alexis-Michel Mugabushaka

[1]*Vicenc.PARISI-BARADAD@ec.europa.eu*
European Research Council Executive Agency, COV 24/161, B-1049 Brussels (Belgium)

Abstract
With the availability of vast collection of research articles on internet, textual analysis is an increasingly important technique in scientometric analysis. While the context in which it is used and the specific algorithms implemented may vary, typically any textual analysis exercise involves intensive pre-processing of input text which includes removing topically uninteresting terms (stop words). In this paper we argue that corpus specific stop words, which take into account the specificities of a collection of texts, improve textual analysis in scientometrics. We describe two relatively simple techniques to generate corpus-specific stop words; stop words lists following a Poisson distribution and keyword adjacency stop words lists. In a case study to extract keywords from scientific abstracts of research project funded by the European Research Council in the domain of Life sciences, we show that a combination of those techniques gives better recall values than standard stop words or any of the two techniques alone. The method we propose can be implemented to obtain stop words lists in an automatic way by using author provided keywords for a set of abstracts. The stop words lists generated can be updated easily by adding new texts to the training corpus.

Conference Topic
Methods and techniques

Introduction
Textual analysis -also referred to as "lexical analysis"," text mining", "co-word analysis" or "linguistic network"- has a long tradition in scientometric analysis. Earlier references can be found in the pioneering work of Eugene Garfield and others (see Garfield, 1967) studying the potential of citation analysis in information retrieval as compared to methods based on terms frequencies. Callon et al. (1983, 1986) introduced the concept of co-word analysis in science and technology studies. This technique was further developed and popularized in scientometrics by the work of Leydesdorff (1989) and researchers at the Center for Science and Technology Studies (CWTS) at the Leiden University (Noyons & van Raan, 1998).

With the availability of vast collections of research articles and better and faster computer tools, which help text analysis, the technique has firmly established itself in scientometric analysis. Nowadays it is used in various contexts: to study the thematic proximity in a collection of documents; to map scientific papers based on concept maps; to detect dynamics and trends of research based, for example, on centrality of concepts or to characterise a particular research community, by identifying relationships between the terms it uses.

While textual analytical techniques differ in degree of complexities and approaches they take, virtually all of them require relatively intensive pre-processing of the input texts. Typically, the following steps are involved in the pre-processing: (1) tokenization, (2) converting to lower case, (3) stemming and (4) removing stop words. For this last step, researchers typically use standard stop words lists obtained from texts in many different domains.

In this paper we argue that using corpus specific stop words might help the textual analysis. The paper is divided in four parts. The next section reviews briefly existing work on stop words and describes in detail two, relatively simple methods, to extract corpus specific stop words. In the subsequent, third, section we present a case study to illustrate the benefits of corpus specific stop words over more general stop words. The concluding remarks discuss limitations and point to future directions.

Related Work

When researchers in scientometrics started using textual analysis, they were standing in long tradition of information retrieval research. Early studies of word frequencies in a text or collection of documents appeared in the last century, when George K. Zip formulated an empirical law that relates terms frequencies (tf) to rank in a frequency ordered word list (Zip, 1932). This frequency characterisation was used later by Hans Luhn to obtain statistical information of words in texts and to compute a relative measure of the significance of individual words and phrases (Luhn, 1958). Using this measure Luhn hypothesized that the most discriminant words are those appearing in the middle of the frequency rank. Salton went a step further by incorporating the document frequency (df) as a measure of the discriminatory capacity of the words (Salton & Young, 1973). They suggested that words can appear in a document collection either in a random manner or concentrated in a few exemplars and they proposed the product of the term frequency times the inverse document frequency (tf • idf) as a measure of the degree of significance: the words appearing in many documents (df high) or with a low presence (tf low) are considered stop words. Based on these frequency descriptions Christopher Fox elaborated in the 90's a list containing stop words (Fox, 1990) extracted from the Brown Corpus of English literature. Although these stop words can be considered the standard or classical list and they have been frequently used, we note two limitations: first they are quite outdated and second they may be too general to take into account the specificities of a collection of texts. They may not be suitable to filter out words belonging to specific research fields or words of recent apparition. As Makrehchi & Kamel (2008) suggest, specific stop words differ from one domain to another.

Several methodologies have been proposed recently to create new stop words lists, customized to particular corpus. Among them, two proposals attracted our attention due to their relative simplicity.

On one hand, an unsupervised method to compute stop words lists arises from the study of the statistical distribution of words, by Church, K. and Gale, W. (1995) and their hypothesis that common stop words follow a Poisson distribution. This has been used to create a stop word list for particular Polish texts (Jungiewicz & Lopuszyński, 2014). We call this approach the *Poisson stoplist.*

Under this hypothesis one assumes that the document frequency of words (df) in a corpus can be estimated (dfe) from their term frequency (tf) and the total number of documents (N) by using the probability theory:

$$\frac{dfe}{N} = 1 - P(0),$$

where $P(0)$ is the probability of not appearing the word. Assuming a Poisson distribution for stop words, the probability of k instances of a word is given by:

$$P(k, \mu) = \frac{e^{-\mu} * \mu^k}{k!},$$

where μ is the average number of instances per document:

$$\mu = \frac{tf}{N}$$

The relation dfe/df is supposed to be close to 1 for randomly distributed terms (stop words) and shows an increase for highly cluttered terms (keywords); although this depends on the corpus, as Jungiewicz and Lopuszyński found when computing their stop word lists for legal texts from the public procurement domain. They realised that their most common stop words had a high variability in their distribution and replaced the Poisson assumption with a negative binomial distribution, which allows a larger variance.

On the other hand, S. Rose et al. (2010) proposed an unsupervised, domain and language independent method to extract keywords from individual texts called RAKE (Rapid Automatic Keyword Extraction) and a supervised method to elaborate stop word lists based on the intuition that words adjacent to keywords tend to be stop words.
RAKE uses stop words to parse the text and extract candidate key phrases (consisting in one or more words). The key phrases are then scored by computing word co-ocurrences and using a metric that favours words belonging to long key phrases. The top T candidates are chosen as keywords (key phrases).
The method proposed by S. Rose to extract stop words from a corpus resorts on accumulating for each word its 'adjacency frequency' (af) and 'keyword frequency' (kf), together with the term frequency (tf) and document frequency (df). Then, given a selection threshold n, the most frequent words with af > kf are chosen as stop words. This method is called by the author *keyword adjacency stoplist* (because it includes primarily words that are adjacent to and not within keywords: Rose et al. 2010, p. 14). We refer to this method as *RAKE stoplist* in this paper.

Case Study: stop list for a collection of abstracts of funded projects

To study the suitability of the above described methodologies and create our own stop words list we applied them to a corpus from abstracts of projects, funded by the European Research Council, in the Life Sciences domain. This corpus consists of 1579 projects covering diverse research areas. The table 1, shows the number of project abstracts by each research area (which corresponds to the scientific panel in which the project was evaluated).

Table 1. Overview of the corpus of abstracts used in the case study

Scientific areas	abstracts	%
Molecular and structural biology and biochemistry	176	11.1
Genetics, genomics, bioinformatics and systems biology	178	11.1
Cellular and developmental biology	164	10.4
Physiology, pathophysiology and endocrinology	176	11.15
Neurosciences and neural disorders	217	13.7
Immunity and infection	168	10.6
Diagnostic tools, therapies and public health	209	13.2
Evolutionary, population and environmental biology	168	10.6
Applied life sciences and biotechnology	115	7.3

Creating stop words

We randomly chose 80% of the abstracts as a training set and the other 20% as a test set.

Following the algorithms outlined in Rose et al. 2010, we wrote a program in Python to create a table (which we call Frequency table) with all the words (12621 in total) of the training set that contains the words, term frequencies (tf), document frequencies (df), keyword frequencies (kf) and adjacent frequencies (af).
This table was used to create both the *Poisson stoplist* and the *RAKE stoplist*. For the later, we set various thresholds to obtain the top n words with the highest term frequency.

Evaluating the stopwords

To evaluate if the corpus-specific stop words improve textual analysis, we use them in extracting keywords. We compare the keywords extracted using those stop words with

author-provided keywords. The idea is that, depending on the stop words used, the keywords extracted will match more or less the ones provided by the authors and the higher the share of matched keywords the better the stop words list.

It should be noted that author-provided keywords do not necessary contain words which also appears in the abstracts. In our corpus, out of 7845 keywords given by the authors only 3494 (44.5 %) where encountered in the abstracts. This means that the precision and F-measure need to be taken into account with care and thus we have not used them for the evaluation of the quality of the stop words list, resorting only to the recall measure, computed as the relation between the total number of correct extracted keywords and the total number of keywords given by the authors, that appear in the abstracts.

We compared the keywords provided by authors with the keywords extracted using the following lists of stop words

1. Standard *Fox stop words list*
2. Stop words list created using the Poisson distribution hypothesis (*Poisson stoplist*)
3. Stop words list computed using keyword adjacency (*RAKE stoplist*)
4. Stop words lists computed using combinations of *Fox,Poisson* and *RAKE*

For keywords extraction we used a Python implementation of the RAKE algorithm (https://github.com/aneesha/RAKE)

1. *Fox stoplist*

This list serves as a baseline for our work and the computation of the recall of the keywords extracted using RAKE algorithm does not need to tune any parameter. The recall obtained is 56.42%.

2. *Poisson stoplist*

To extract the stop words using this approach we need first to set the threshold for the relation dfe/df. To do that we computed the mean and standard deviation of the dfe/df for all the *Fox* stop words that appear in the training set. Figure 1 shows the plot of these values, where the mean (dfe/df) + std(dfe/df) is 1.55. There are only 14 Fox stop words excluded from the list and apart from the words (*ordering, right and small*) their term frequency is very low. We have used this threshold to obtain the stop words from our training data appearing in at least 10 documents (df>10) and we have obtained a list of 2008 words that gives a recall of 58.25% in our test set, which is better than the *Fox stoplist*.

3. *RAKE stoplist*

To use the RAKE approach we extracted all the words from the training set with af>kf and created an ordered table, sorted in descending order of word occurrence (tf). This table consisted in a list of 2045 candidate stop words. To choose the top best frequency rank we tested subsets of these lists and computed their recall values. The result obtained using all the words in the list was 45.42 % of recall and the results improved by removing words from the list, having a peak at a 53.31% of recall, when using the first 185 words of the rank.

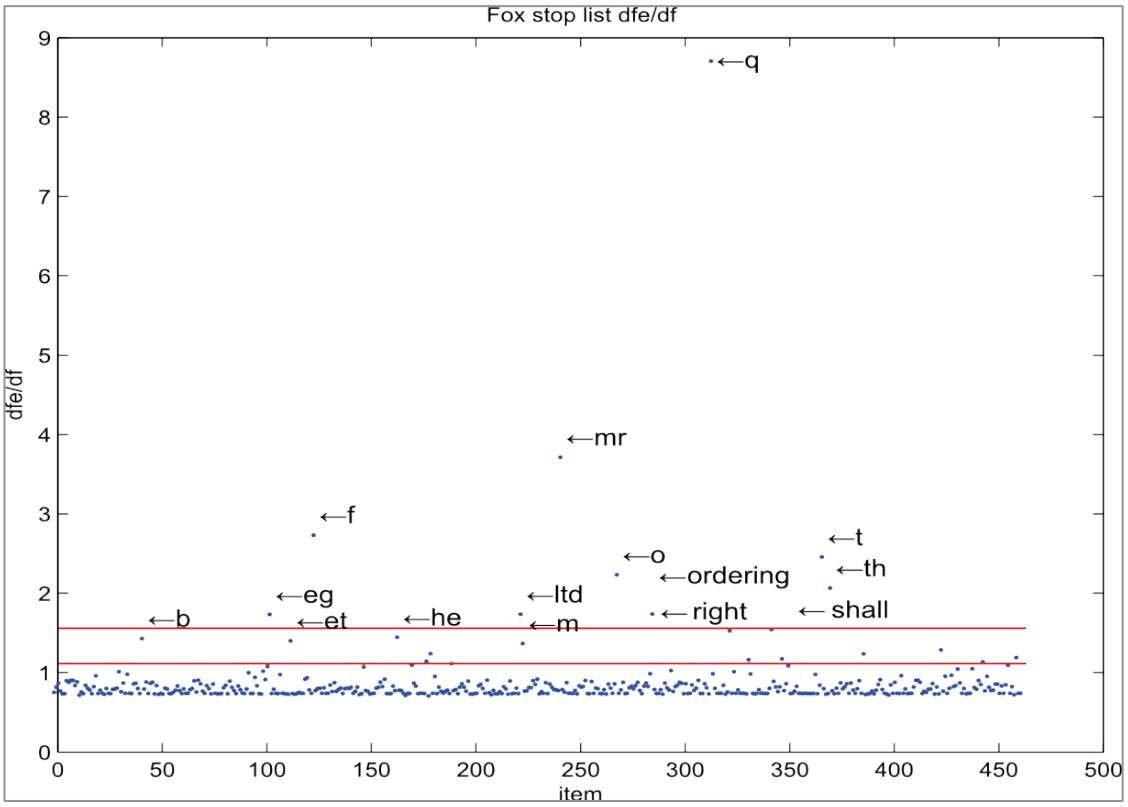


Figure 1. Fox stoplist dfe/df values found in the training corpus. Just a few words are above the standard deviation limit, and they are rarely found (tf very low).

4. *Combinations Poisson and RAKE*

Since the *RAKE stoplist* gave us worse results than the *Fox stoplist*, we tried to combine them with the Poisson approach (*RAKE-Poisson*) and we extracted the words with df>10,af>kf and dfe/df<1.55. This improved the previous results, giving a recall of 62.34%. Note that the condition dfe/df> r can also be seen as an adaptive threshold on tf, since, under the Poisson distribution, it can also be expressed as:

$$tf > N * \ln\left(\left(\frac{df}{N}\right)r - 1\right),$$

and instead of choosing a minimum common tf for all the words, we adapt the tf to each word's df. In Figure 2 we have plotted the df of the RAKE stop words (tf>0) , together with the Fox stop words found in the *RAKE stoplist*. Also we plotted the dfe/1.55 curve which shows the limit above which the words belong to *RAKE-Poisson stoplist*.

After inspecting the frequencies of the RAKE-Poisson stop words we found words expected to appear in Life Sciences texts and we questioned ourselves if their removal from the stoplist would improve the recall results. To check it we removed them by hand and the recall increased to 64.56 %. A more detailed inspection of the stoplist frequencies allowed us to see that just a few words (6 in total) belong to the life sciences domain (genetic, disease, protein, molecular, gene, cell), all them with kf>60 had a 1.1<dfe/df <1.55. In all them the af/kf relation was less than 5 (af/kf<5). This data gave us the intuition that we needed to decrease the dfe/df threshold and also to be more strict on the af/kf condition, so we tested a stoplist consisting in the RAKE intersection with Poisson stoplist (df>10 and af>5*kf and dfe/df<1.2) which gave a recall of 68.69%, being this the best result. We call it the *RAKEm-Poisson stoplist*.

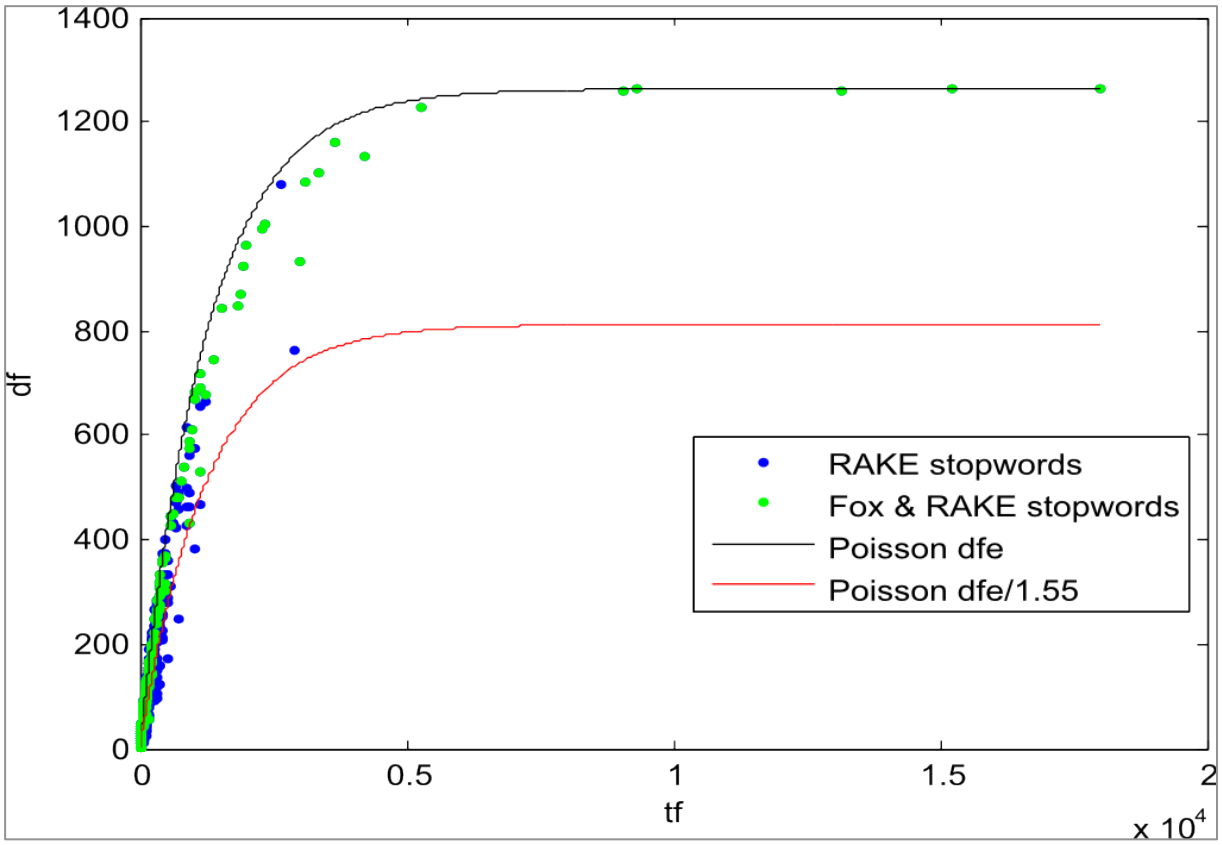


Figure 2. RAKE and Fox stop words. We can see that the Fox stop words follow the Poisson distribution better than the RAKE stop words, which appear more concentrated at low df values.

Conclusion

Our aim is to obtain stop words that help to provide meaningful and significant keywords that summarize the texts; the validation of the stoplists we did was based using the author given key phrases which most of the times had fewer words than the ones obtained using RAKE. We think that this circumstance is favouring standard stoplists since they will still produce single word keywords given by authors and end up yielding overall recall values similar to specific domain stoplists. Therefore we plan as a future work to use measures that evaluate semantic value of the key phrases.

We would like to remark that the RAKE-Poisson stoplist can be obtained from the word frequencies and the author keywords, without further human intervention. Our future work involves also the automatization of the computation of the best af/kf and dfe/df thresholds to generate the *RAKEm-Poisson stoplists*.

References

Blanchard, A. (2007). Understanding and customizing stopword lists for enhanced patent mapping, *World Patent Information, 29*(4), 308-316.

Callon, M., Courtial, J.-P., Turner, W. A., & Bauin, S. (1983). From translations to problematic networks: an introduction to co-word analysis. *Social Science Information, 22*, 191-235.

Callon, M., Law, J., & Rip, A. (Eds.). (1986). *Mapping the Dynamics of Science and Technology*. London: Macmillan.

Church, K. and Gale, W. (1995). Poisson mixtures. *Journal of Natural Language Engineering*, *2*, 163-190.

Fox, C. (1990). A stop list for general text. *ACM-SIGIR Forum*, *24*, 19-35.

Garfield, E. (1967). Primordial concepts, citation indexing and eIistorio-bibliography. *The Journal of Library History, 2*(3), 235-249. http://www.garfield.library.upenn.edu/essays/v6p518y1983.pdf

Jungiewicz, M. & Lopuszyński, M. (2014). Unsupervised keyword extraction from Polish legal texts. *Advances in Natural Language Processing*, 65–70. Springer LNCS.

Leydesdorff, L. (1989). Words and co-words as indicators of intellectual organization. *Research Policy*, 18(4), 209-223.

Luhn, P. (1958). The automatic creation of literature abstracts, *IBM Journal of Research and Development, 2*(2) 159–165

Makrehchi, M. & Kamel, M. (2008) *Automatic Extraction of Domain-specific Stopwords from Labeled Documents*. Berlin / Heidelberg: Springer.

Noyons E. C. M. & Raan A. F. J. (1998). Monitoring scientific developments from a dynamic perspective: Self-organized structuring to map neural network research. *Journal of the American Society for Information Science*, *49*(1), 68–81.

Rose, S., Engel, D., Cramer, N. & W. Cowley, W. (2010). Automatic keyword extraction from individual documents. *Text Mining: Applications and Theory*. John Wiley & Sons, Ltd.

Salton, G. & Yang, S. (1973). On the specification of term values in automatic indexing, *Journal of Documentation,* 29(4), 351–372.

Sinka, M.P. & Corne, D.W. (2003). Towards modernised and web-specific stoplists for web document analysis, *IEEE/WIC International Conference on Web Intelligence,* 396-402.

Zipf, K. (1932). S*elective Studies and the Principle of Relative Frequency in Language*. Cambridge: MIT Press.