Research Evaluation **23** (2014) pp. 261–271
Advance Access published on 2 July 2014

doi:10.1093/reseval/rvu014

The long-term influence of collaboration on citation patterns

Ali Gazni[1,3,*] and Mike Thelwall[2]

[1]Vice President in Research Affairs, ISC, Shiraz, Iran, [2]Statistical Cybermetrics Research Group, University of Wolverhampton, Wulfruna Street, Wolverhampton WV1 1LY, UK and [3]Member of Research Department of ISC, Regional Information Center for Science and Technology, Shiraz, Iran

**Corresponding author. Email: ali.gazni@isc.gov.ir*

This study assesses the long-term impact of collaboration in terms of the extent to which collaborators cite each other's works and cite the same publications as each other. The results are based on coauthorship of academic articles during 1990–2010. Although the number of citations to, and common references with, collaborators both increase as the number of collaborators increases over time, these differ between collaborators. For example, many authors do not cite their collaborators and many collaborators do not cite any of the same references as each other. In contrast, many authors cite their collaborators extensively and many collaborators have many of the same references as each other. The extent of citing collaborators and citing the same references as cited by collaborators varies with the impact of the collaborators. These widely different properties may reflect some collaborators working in completely different research areas, others working in the same broad research area, and still others working within a narrow research area. Alternatively, some collaborators may learn from or monitor each other while others do not.

Keywords: scientific collaboration; collaboration strategies; collaborative networks; organization of science.

1. Introduction

Scientific collaboration is encouraged by policymakers within and across countries by providing funding to create and sustain scientific networks (Porac et al. 2004; Defazio, Lockett, & Wright, 2009). This has contributed to the increasing dominance of teams in scientific production (Adams, 2012; Wuchty, Jones, and Uzzi, 2007). Collaborative work may help scholars to get results that are more reliable because more than one person is involved (Thagard, 1997). Social connections could also reduce competition, increase trust, facilitate the exchange of complex knowledge, create social support for the adaption of a piece of knowledge, and help to speed up knowledge creation and innovation (Ahuja, 2000; Reagans and McEvily, 2003; Fleming, Mingo, and Chen, 2007). Thus, despite some criticism (e.g. Katz and Martin, 1997: 1), scientific collaboration is generally perceived as being beneficial (e.g. Adams, 2012: 1).

Scientists may collaborate with each other because of complementary scientific expertise, political factors, socioeconomic factors, resource accessibility, and social networks. The collaboration itself may then facilitate knowledge exchange, transfer, sharing and use based on common needs, languages, understanding, goals, and tasks through formal and informal interaction among collaborators during their academic lifetimes (Schrage, 1995; Katz and Martin, 1997; Jassawalla and Sashittal, 1998; Lambert, 2003; Sonnenwald, 2007). Nevertheless, there is no empirical evidence about the long-term effects of collaboration on individual scholars.

The main aim of this study is to gain insights into the long-term impact of scientific collaboration by measuring the extent to which collaborators cite each other and reference the same documents. Although citations are only one aspect of scientific activity and do not necessarily indicate knowledge sharing or use, they are a convenient source of data to investigate one aspect of scientific collaboration on a large scale (e.g. Yan and Sugimoto, 2011). In addition, citations reflect knowledge flows between collaborators to some extent, and therefore citation patterns may give indirect evidence of knowledge flows associated with collaboration.

2. Coauthorship, information behavior and knowledge flows

When researchers collaborate, they presumably learn from each other to some extent. Even if they work on completely separate tasks and do not extensively discuss issues or teach each other theories or methods then they can at least be expected to read each other's contributions to their joint publications. Thus, it seems likely that collaborations may generate knowledge flows between collaborators. In this context, learning refers to the exchange of knowledge, and collaboration is one of the most valuable channels for this (Lambert, 2003). Intuitively, common work needs a shared language and a common contextual understanding and so collaboration can help in the integration of knowledge, efforts and capabilities, especially in interdisciplinary research (Maglaughlin and Sonnenwald, 2005; Andrade, de Los Reyes, and Martín, 2009). However, Katz and Martin (1997) differentiate between different types of collaborators, from those who just give general advice to those who actively participate in the research, either practically or theoretically. Nevertheless, coauthorship is widely accepted as a proxy for collaboration, when it is measured, although not all collaborations are rewarded with coauthorship (Laudel, 2002) and vice versa.

Simultaneously analyzing coauthorship and citation networks for the same researchers can help to reveal the influence of collaborators on knowledge use in scientific networks. However, the assessment of this influence in social networks remains a challenge (Bakshy et al. 2012). If author A cites author B's paper, it suggests that A perceives B's paper as containing relevant knowledge. If this is the case then this can also be considered as *knowledge flow* from B to A via their papers. Similarly, if A and B cite the same sources of knowledge then they may have similar *information needs* or similar *patterns of information use* or may have *acquired the same knowledge*. Thus, many identical citations suggests that there may be many common information needs, uses or knowledge, and this may also indicate shared understandings and even a close intellectual relationship.

Although knowledge flows are invisible, there have been attempts to measure them indirectly to build evidence-based models. Any such measurement is necessarily based upon a set of context-specific assumptions (Krugman, 1991). Authors' ideas can be published in journals and diffused through citations across scientific communities in terms of journals, fields, countries, and institutions (Liu and Rousseau, 2013). These notions encompass both creation and dissemination in science (Soler, 2007). For instance, citations have been used to indicate knowledge flow across disciplines as well as the speed at which knowledge flows between fields (Rinia et al. 2001, 2002; Yan et al. 2013). Moreover, the transfer of codified knowledge from the public research sector to the private sector has also been measured by analyzing citations from industrial research publications to public research publications (Tijssen and Van Leeuwen, 2006).

Bibliometric methods have previously been used to map research specialties and to determine the intellectual and social structure of academic communities (Morris and Van der Veer Martens, 2008). Citation analysis has also been used to study scholars' information use (Wilson, 1996; Burright, Hahn, and Antonisse, 2005; Kuruppu and Moore, 2008; Bigdeli and Gazni, 2012; Bigdeli et al. 2012). Information seeking begins when scholars perceive a need for information. They then make a demand on a formal (e.g. digital library) or an informal (e.g. colleagues or collaborators) channel to find relevant information, which they may then use (Wilson, 2006) and perhaps also cite. Many years ago, it was noticed that some information comes from invisible colleges (Crane, 1969; Griffith, Jahn, and Miller, 1971), which suggests that there may often also be information flows among collaborators (Price, 1986).

3. Citation and scientific collaboration

Citations can be a byproduct of information use, as discussed above, and therefore citation analysis as a method could shed light on a user's past and present information behaviors (Smith, 1981). In comparison to other methods, citation analysis can provide easy and efficient access to large-scale (partial) evidence about information uses, needs, and behaviors (Fuchs et al. 2006). Nevertheless, informal communication between two scholars may exist for a long time before they become coauthors and authors may continue to collaborate informally after a coauthorship. They may also follow each other's works and exchange preprints, which makes it difficult to determine the influence of collaboration on coauthored papers alone.

Most studies of the relationship between collaboration and citation for authors, institutions, and countries have found that having more authors on a paper associates with more citations (Didegah and Thelwall, 2013). This relationship varies depending on the countries, states, and regions investigated (Levitt and Thelwall, 2010). In some studies, collaboration seems to result in higher quality science or more connections to other scholars (Goldfinch, Dale, and DeRouen, 2003; Franceschet and Costantini, 2010; Gazni and Didegah, 2011).

The extent to which researchers reference current collaborators varies considerably between disciplines, from under 1% of history references to ~25% of astrophysics and astronomy references (Wallace, Larivière and Gingras, 2012). This article, the most similar to the current paper, analyzed eight Web of Science categories (astronomy and astrophysics, atmospheric science and meteorology, biochemistry and molecular biology, economics, history, neurology and neurosurgery, organic chemistry, and sociology), and did not perform author name disambiguation, arguing that it was likely to affect only ~5% of author names and hence should not have a major impact on the results. They did not study the number of common references with collaborators and did not investigate differences between authors with different levels of impact. They also focused on eight subject fields from the natural, medical, and social sciences and humanities, whereas the current study analyses science as a whole.

From a different perspective, more collaboration between two institutions increases the probability of citations between them, at least for 59 library and information science journals. Similarly, more citations between institutions increase the chance of future collaborations for articles in these journals (Yan and Ding, 2012). Moreover, patent collaborations increase the probability of citations and, as the path length increases (comparing level 1 to level 2 coauthors), this probability steadily declines (Singh, 2005). More generally, Johnson and Oppenheim (2007) found a positive relationship between social closeness and number of citations by investigating 16 individuals using citation analysis and a questionnaire. Nevertheless, many citations made by information scientists refer to authors outside of their immediate social connections.

4. Research questions

Although scientific collaboration seems to be widely recognized as important to most types of science, this phenomenon needs a deeper understanding of the influence of collaborators on each other during their academic lifetime rather than just for individual projects. An empirical study of information flows in collaboration networks could give more insights into the extent to which knowledge is exchanged between collaborators. It can also help to develop an understanding of scholars' information behavior through coauthorship and citation networks. Information about how different scholars form their social and informational environment with respect to their collaboration networks could also give a better understanding of the structure and social structure of science.

No previous studies have determined the extent to which authors cite their collaborators in the long term (i.e. longer than 5 years). Moreover, no studies have analyzed the extent to which authors cite the same references as their collaborators, in either the short term or the long term. This research seeks to answer the following questions:

(1) To what extent do authors cite the same publications as those cited by their coauthors in the long term?
(2) To what extent do authors cite their coauthors in the long term?
(3) Do the answers to questions 1 and 2 change over time?

5. Data and methods

5.1 Data

All Thomson Reuters Web of Science (WoS) citable documents, including articles, reviews, and proceedings papers indexed during 1990–2010 were extracted. This 21-year period is recent but old enough to allow articles to attract citations, and covers a substantial number of years. The exact start and end years are relatively arbitrary but were chosen to be round numbers. The references in these documents to articles in journals indexed in WoS during 1990–2010 were also extracted. The WoS documents contained 504,450,188 references, 50% of which were citations in WoS journals from 1990 to 2010. WoS was used rather than Scopus because of the length of its coverage and also because WoS journals are widely perceived as being the most prestigious (Brody, 1995; Ohniwa et al. 2004; Kurmis and Kurmis 2006) although most are from only a few publishers (Didegah and Gazni, 2011).

5.2 Disambiguating author names

In WoS, one author may have multiple names and one name may belong to several authors (Reuther and Walter, 2006). When investigating authorship, the accuracy of the results can therefore be improved by conflating multiple names from a single author and disambiguating one name that belongs to different persons. Different techniques to measure the similarity of two strings could potentially be used for recognizing multiple names of one author, such as edit distance, Soundex, and Jaro-Winkler distance, which may be able to help identify minor variations in author names. Another way to identify this problem is to match only the last names and the initial letters of first and middle names, known as the blocking mechanism. This method dramatically reduces the computational cost of identifying author name variants. Moreover, blocking by last name and first name initial results in a loss of only ~2% of the possible variants and so has a high degree of recall (Bilenko, Kamath, and Mooney, 2006; Smalheiser and Torvik, 2009).

The two above mentioned issues can also be addressed with additional available information, such as email addresses, affiliations, self-citations, topics, coauthors, and research domains, all of which could be combined with other evidence, such as commonality and rarity of terms or information (Torvik et al. 2005; McRae-Spencer and Shadbolt, 2006; Culotta et al. 2007; Kanani and McCallum, 2007; Song et al. 2007; Kang et al. 2009). Combining these features can give more accurate results. For example, Kang et al. (2009) clustered articles based on coauthor networks, noting that more than 85% of author ambiguities were resolved by coauthorship. McRae-Spencer and Shadbolt (2006) focused mainly on self-citation and coauthorship to cluster papers, achieving a precision of 0.997 and a recall of 0.818 for determining whether one name belongs to several authors. Self-citation is perhaps the most important single feature for author name disambiguation and can be easily calculated (Levin et al. 2012).

For disambiguation, author names were first broken into blocks consisting of full last names, first name initials, and middle name initials, if available, giving 5,508,223 blocks. Based on the above research, less than 2% of authors should be split between blocks, which seems low enough to be ignored, but individual blocks may represent multiple authors, which is likely to be a more serious problem.

To split up blocks representing different authors, self-citations were used as follows. An article self-citation network was constructed by connecting two articles in the network whenever a reference in one of the papers cited another paper, with both papers apparently having the same author (i.e. with authors approximated by blocks, as described above). All articles in this network were then clustered with a modularity optimization clustering algorithm that has previously been applied to citation networks and coauthorship networks (Blondel et al. 2008; Lambiotte and Panzarasa, 2009; Wallace, Gingras, and Duhon, 2009; Zhang et al. 2010). This method was chosen rather than other well-known clustering algorithms, as the number and size of the clusters produced with this method have the maximum modularity property that seems reasonable for author name disambiguation. The clustering produced 669,217 clusters of articles including 11,890,170 of the 17,981,346 citable documents, with the remaining articles being disconnected. Blocks occurring in different clusters of articles were then split to create a new 'author' for each cluster. High precision was expected for the authors disambiguated in this way (i.e. most articles clustered together should be from the same person) because authors are more likely to cite their own research rather than the research of another person with the same name. Nevertheless, the method may incorrectly split authors that operate on separate topics without cross-citing themselves in them and may fail to split authors with common names (e.g. W. Zhang) who are reasonably likely to cite other authors with the same name. This process resulted in 5,557,941 disambiguated authors and these were used for the sampling. The same disambiguation was used for the coauthors of these people, resulting in 4,686,209 disambiguated collaborators.

Authors' e-mail addresses were used to estimate the precision of the author disambiguation process because e-mails have close to 100% precision, when present, but not 100% recall because an author may have multiple e-mail addresses or may change their e-mail address (Levin et al. 2012). WoS does not clearly match e-mail addresses to author names, however, which is a problem for multi-authored articles. Different authors use different format for their e-mail addresses, such as full first and/or middle name and/or last name, dot and/or hyphen, letters or numbers. To accurately match e-mail addresses to people, we matched only those addresses that start with the author's first name initial and middle name initial, if available, followed by a dot and the author's full last name (e.g. w.zhang@) for the test collection. Only authors with an e-mail address assigned to them in all of their papers were matched for the precision test. There were 25,057 blocks in which all authors had an e-mail address assigned to them. None of these blocks contained multiple e-mail addresses, but 3,683 blocks contained e-mail addresses that occurred in at least one other block. Hence, 21,374 of the 25,057 blocks (85.3%) were completely correct in the sense of containing only one e-mail address and that e-mail address never occurring in another block. This level of accuracy is much lower than reported in previous studies, which may be because of the much larger data set analyzed and the much longer time period covered. Nevertheless, it is likely to be an underestimate of the true accuracy because the larger blocks are probably the most problematic and these tend not to be included in the test set because at least one e-mail address is likely to be missing from them.

To test the extent to which the author disambiguation accuracy may affect the results, the percentage of (1) citations to and (2) shared references with collaborators were computed for all authors using the same 25,057 blocks from the test set described immediately above. There were 11.8% citations and 48% shared references. After eliminating the 3,683 incorrect blocks, there were 12.6% citations and 51.3% shared references, suggesting that, on the full data set, citations may be underestimated by ~0.8% and shared references may be underestimated by ~3.3%. For the complete set of 38,152 sampled authors, there were 10.6% citations and 44.7% shared references. The above logic suggests that the true figures may be closer to 10.6% + 0.8% = 11.4% and 44.7% + 3.3% = 48.0% although, as the values for the full data set are different from those from the test set, the test values are not reliable and the true figures may be more different than this.

5.3 Determining impact classes

Thomson Reuters' Essential Science Indicators (ESI) records the top percentile of papers based on their citations within their scientific field: 0.01, 0.10, 1, 10, 20 and 50%. These were used to classify authors according to the number of normalized citations received, as described below. A 100% class and an uncited class were added to the ESI classification because it does not cover all papers.

Time and discipline both affect the number of citations received by a paper. Hence, it is important to normalize the number of citations received by each article before conducting any comparisons. To normalize the number of citations received, each journal was assigned to one of 22 subject fields using the journal list in the Thomson Reuters ScienceWatch Web site. The average number of citations received by all papers in each subject field and in each year was then computed. Finally, the number of citations received by each paper was divided by the average number of citations for its subject field in the paper's publication year (Adams, Jackson, and Marshall, 2007).

To make the processing manageable, a random sample of 38,152of the 5,557,941 disambiguated authors was selected. This sample size was calculated to be large enough for an accuracy of 0.5% for a 95% confidence interval of the proportion of authors with any given property. Around 402,452 of the 17,981,346 citable WoS documents from 1990–2010 belonged to these authors. These were also the corresponding authors of 152,020 citable documents, and their collaborators produced 4,465,059 citable documents in the same period.

5.4 Calculations

To answer the research questions, references in the papers of the 38,152 sampled authors in which they were the corresponding author (152,020 papers) were analyzed. The corresponding author has the most important role in a paper across all fields, and presumably tends to have the main responsibility for a paper's references.

Author name order in a paper could help to indicate the type of contribution that they have made to the paper. For example, the first and the last author may have the most important roles, depending on the fields (Mattsson, Sundberg, and Laget, 2011). For instance, in biomedical science, the first author has typically conducted the experimental work whereas the last author has supervised it (Stokes and Hartley, 1989). Similarly, in molecular biology, the first author has probably done the main work, whereas the last one probably led the project (Herbertz and Muller-Hill, 1995). In contrast, in sociology, the last author probably made the smallest contribution to the paper (Moya-Anegón et al. 2013). The first author usually had the main responsibility in the research project and did most of the work in medical articles (Riesenberg and Lundberg, 1990). Based on 35 years of papers from the top five most cited institutions in the world, 81% of first authors and 16% of last authors are corresponding authors (Bigdeli and Gazni, 2012).

To calculate the number of citations that an author has in common with their collaborators (question 1), references in the author's papers (for which they were the corresponding author) were compared with references in their collaborators' papers (whether or not they were the corresponding author) and the number of common citations was counted. Papers coauthored between the two authors were excluded from both data sets. To investigate an author's citations to their collaborators (question 2), references in the author's papers (for which they were the corresponding author) were checked to determine the number citing their collaborators' papers whether or not they were the corresponding author. This method excludes citations to and from papers coauthored between the two authors.

To estimate the growth in the number of references that an author has in common with those of their collaborators and the growth in the number of citations to an author's collaborators over time (question 3), the publishing span of each author was estimated based on their first and the last publication years. The number of citations to their collaborators' works and the number of citations in common with their collaborators during their publishing span were computed separately for each author. The growth in the number of collaborators for each author over time was also calculated based on their publishing span.

5.5 Collaborators

In this study, an author's collaborators were defined to be their coauthors, considering only articles, reviews, and proceedings papers in which the author was a coauthor. For each author, all collaborators and their papers were used for estimating common references with, and citations to, collaborators, except for determining the number of collaborators in each year for an author (Fig. 4).

Table 1 shows an example of collaborations in the publications of an author. An author's collaborators could be counted in different ways. If B coauthors 10 papers over three years with A, then B counts as three 'collaborators' of A, with 10 joint papers (Table 2, method A). Here, each coauthor counts only once in each year irrespective of the number of coauthored papers in that year. This method is a compromise between (1) analyzing each *collaboration* separately (Table 2, method B) and (2) analyzing each *collaborator* separately (Table 2, method C). The latter would be the natural method, but has the disadvantage that occasional collaborators are treated equivalently to long-term collaborators. The compromise method of counting the number of collaborators in each year was chosen because it gives a higher weighting to long-term collaborators; this method is only used for Fig. 4.

6. Results

6.1 To what extent do authors cite the same publications as those cited by their collaborators?

On average, 44.7% of the references in an author's papers on which they were the corresponding author could also be

Table 1. An example of collaboration

Article	Publication date	Collaborators
1	2000	A, B, C
2	2000	A, E
3	2001	A, B
4	2002	A, B, E
5	2002	A, F, G
6	2002	A, C
7	2003	A, M

Table 2. Different ways of calculating the number of collaborators

Years	Number of unique collaborators (Method A)	Number of unique collaborators (Method C)	Number of collaborators per paper (Method B)
2000	4	4	5
2001	2	0	2
2002	6	2	8
2003	2	1	2

found in the references of their collaborators in those papers where the given author was not a coauthor, whether or not they were the corresponding author. On average, each author has at least one common reference with 70% of their collaborators. Counting the references that an author has in common with all of their collaborators, on average, 31.5% are also referenced by one or both of two of their collaborators (core zone), 35.9% of the references are also referenced by one or more out of six other collaborators (middle zone), and the remaining 32.6% are also referenced by at least one of the remaining collaborators (peripheral zone).

Author impact classes were determined from the normalized total citations received by them. The proportion of references in common with collaborators out of all of an author's references classes was calculated for each impact class. Generally, high-impact authors had more shared references with collaborators than did lower-impact authors. From Fig. 1, from the top 0.01% of authors to uncited authors, the percentage of common references with collaborators decreases gradually, a linear relationship.

The authors and their collaborators were labeled into different impact classes based on their normalized total number of citations received (Table 3). High-impact authors have more common references with each other than they have with lower-impact authors and vice versa (Table 3), suggesting that they may be working in different areas or with different parts of the academic literature (e.g. high-impact international journals vs. low-impact national journals or a high citation specialism vs. a low citation specialism). Except for un-cited authors that demonstrate irregular patterns compared with authors in the other impact classes, the percentage of common references with the top 0.01–1% collaborators gradually decreases from high-impact to lower-impact authors (Table 3). All groups of authors, including the uncited authors, have the lowest percentage of common references with the uncited authors.

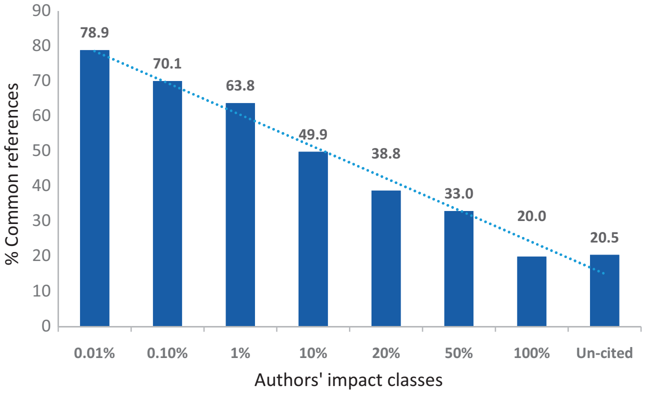

Figure 1. Authors common references with collaborators with respect to the authors' impact classes.

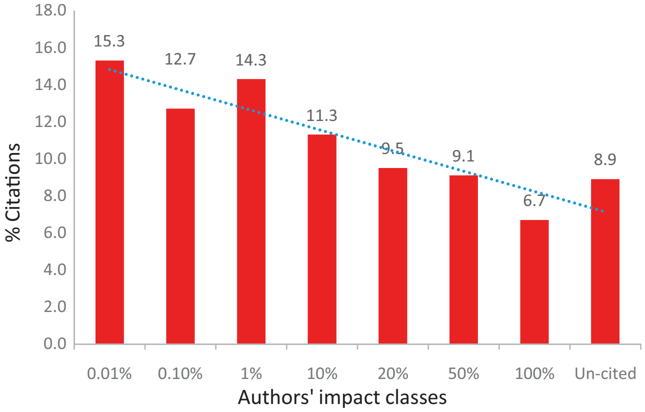

Figure 2. Authors' citations to collaborators with respect to the authors' impact classes.

6.2 To what extent do authors cite their coauthors?

Of the 38,152 disambiguated authors and their WoS collaborators, on average, 10.6% of all their citations in corresponding author papers referred to collaborators' papers where they are not coauthors. On average, the authors cited 12.5% of their coauthors' WoS papers that were not coauthored with them. Counting the citations of an author to their collaborators' papers showed that 38.7% of these were citations to just two collaborators (core zone), 28.9% are citations to four other collaborators (middle zone), and the remaining 32.4% are citations to the rest of the collaborators (peripheral zone).

Authors' normalized total citations received were used to categorize them into different impact classes and the

Table 3. Authors' common references with collaborators (percentages) based on collaborators' and authors' citations impact classes

Collaborators				Authors				
Authors' citation impact class	Top 0.01%	Top 0.1%	Top 1%	Top 10%	Top 20%	Top 50%	Top 100%	Uncited
Top 0.01%	11	7	2	0.7	0.28	0.2	0.3	2
Top 0.1%	32	26	11	5	2.54	2	2	3
Top 1%	37	35	30	21	16.38	14	11	18
Top 10%	16	24	40	47	44.24	39	33	44
Top 20%	2	4	9	13	16.95	17	14	10
Top 50%	2	3	7	11	15.38	21	23	15
Top 100%	0.3	0.7	1	3	4.03	6	17	7
Uncited	0.03	0.03	0.1	0.2	0.20	0.3	0.5	0.8

Table 4. Authors' citations to collaborators' papers (percentages) based on collaborators' and authors' citations impact classes

Collaborators				Authors				
Authors' citation impact class	Top 0.01%	Top 0.1%	Top 1%	Top 10%	Top 20%	Top 50%	Top 100%	Uncited
Top 0.01%	15	10	4.3	2.1	0. 8	0.4	0.6	5
Top 0.1%	39	25	18	10	5	4	3	10
Top 1%	35	38	43	37	30	27	19	21
Top 10%	10	23	29	41	47	45	41	46
Top 20%	0.4	2	3.3	6.0	10	12	13	9
Top 50%	0.2	1	1.6	3.5	6	10	15	8
Top 100%	0.02	0.1	0.15	0.5	1	2	7	2
Uncited	0	0	0	0	0	0	0	0

citations to collaborators' papers were counted where they were not coauthors. High-impact authors have a tendency to cite their collaborators compared with lower-impact authors (Fig. 2). This figure also shows the existence of an approximately linear relationship between an author's impact class and their citations to collaborators.

Patterns of authors' citations to their collaborators were counted with respect to their impact levels (Table 4). Highly cited authors tended to cite high-impact collaborators and less cited authors cite lower -impact collaborators more than high-impact authors (Table 4). Generally, except for un-cited authors, the percentage of citations to the top 0.01–1% collaborators decreases from the top 0.01% authors to the top 100% authors. Conversely, the percentage of citations to the top 20–100% collaborators gradually increases from the top 0.01% authors to the top 100% authors.

6.3 Do authors' citations to their collaborators increase over time?

In general, the authors increased their number of collaborators, citations to collaborators' papers, and references in common with collaborators linearly over time (Figs 3 and 4). The different team sizes and the average number of papers per author in the different fields, the field sizes, and hyper-authorship are some factors that could affect the average number of collaborators across all fields. On average, in their first publishing year, an author has five different collaborators. On average, and over 21 years, the average number of different collaborators includes 23 person names. Similarly, the average number of references in common with collaborators starts from eight references in the first year of publishing and reaches 28 in the 21st year. Moreover, the number of citations to collaborators increases from three per collaborator to nine per collaborator during 21 years of publishing.

Authors were categorized into different impact classes, and their common references with (Fig. 6) and citations to their collaborators (Fig. 5) were calculated over time, considering only papers for they were not coauthors. The top 0.01% of the authors and uncited authors were eliminated from these two figures, because the top authors are different from the other groups and this made the figure complex, and the uncited authors are typically not in the scientific network for long (median = 1). Figure 6 shows that the number of citations to collaborators increases over time in all impact classes, but at different rates, especially for the top 0.1, 1 and 10%, with R^2 values of 0.88, 0.93, and 0.77 for a linear relationship, respectively. The top 0.1% of the authors have about three citations to collaborators in the first year and this number increases to around 25 citations over 21 years. Similarly, the top 1%

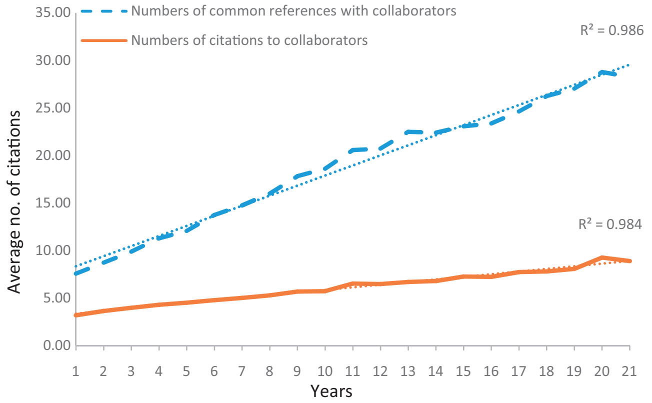

Figure 3. Common references with collaborators and citations to their works over a 21-year period.

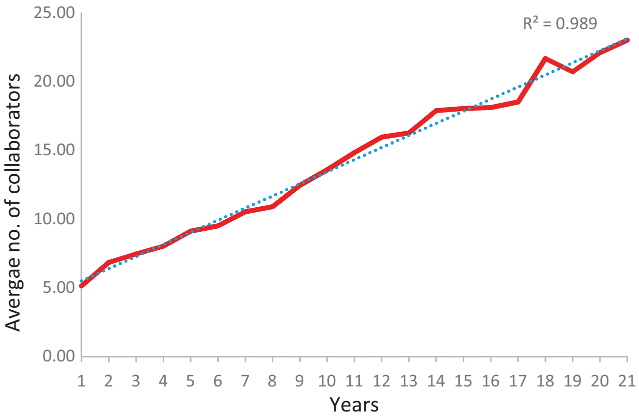

Figure 4. Average number of collaborators of authors over a 21-year period.

of the authors start with about five and end with about 16 citations to collaborators and the top 10% of the authors start with about four and end with about seven. The top 20, 50, and 100% of authors have lower rates of change over time, with R^2 values of 0.1, 0.04, 0.22 for a linear relationship, respectively, and have less citations to collaborators compared with the first groups of authors. Figure 5 reveals that the number of common references with collaborators increases gradually, especially for the top 0.1, 1, and 10% of authors with R^2 values 0.89, 0.89, and 0.77, respectively, for a linear relationship. The top 0.1% of authors has the highest growth rate over time, which started with about seven common references with collaborators and ended with around 111 common references. The top 1% of authors has 10 common references with collaborators in the first year and this number increases to 57 after 21 years. The top 10% of authors increase from 9 common references with collaborators to 23 common references. The top 50 and 100% of authors have again a lower rate of change in common references with collaborators over time compared with the other group of authors.

7. Discussion

The results could be the product of knowledge flows resulting from coauthorship in the sense of authors citing collaborators or citing the same references as do their collaborators because of one learning from or monitoring the other. Alternatively, the results could mainly reflect the structure of the fields in which the collaborators operate, with authors citing collaborators or citing the same references as do their collaborators because they are working in the same field. Hence, it is not possible to state that the collaboration itself has caused the patterns found in this article and further research would be needed to assess whether this is the case.

If the citation and reference results primarily reflect knowledge flows between collaborators then the results suggest that collaboration often does not lead to knowledge flows, because authors have different level of citations to and shared references with different collaborators, a core-periphery structure. Moreover, many collaborators have no any citations to and common references with each other. Perhaps an author's collaborators range from those who have many common needs, understanding, goals, and tasks to those who have little in common. This partly supports previous studies claiming that collaborators increase knowledge flows, facilitating the exchange of complex knowledge, accelerating knowledge creation and innovation.

Following the answers to the first question, an author's collaborators could be divided into three classes:

- a small group of collaborators with a high degree of common reference use with the author, probably indicating shared membership of a narrow research area;
- a large group of collaborators with a few shared references with the author, suggesting membership of a common wide research area or shared methods or theoretical underpinnings;
- a large group of collaborators with no references in common with the author, suggesting that these collaborators either publish no papers that are not collaborative with the author or that they work in a completely different research area and have little or no theory or methods in common.
- Similarly, the answers to the second research question suggest the existence of the following.
- a small group of collaborators that are extensively cited by the author, probably indicating shared membership of a narrow research area or a deep dependence on the collaborator for theory or methods;
- a large group of collaborators that are occasionally cited by the author, suggesting membership of a common wide research area or occasional use of the collaborator's theory or methods;
- a large group of collaborators with no references in common with the author, suggesting that these collaborators either publish no papers that are not

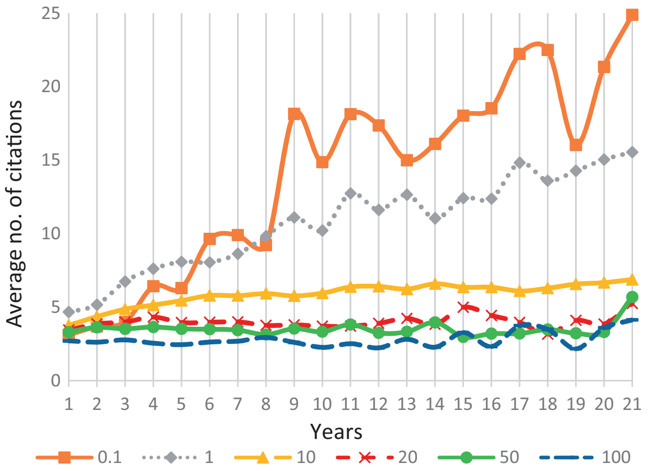

Figure 5. Citations to authors over 21 years based on the authors' impact classes.

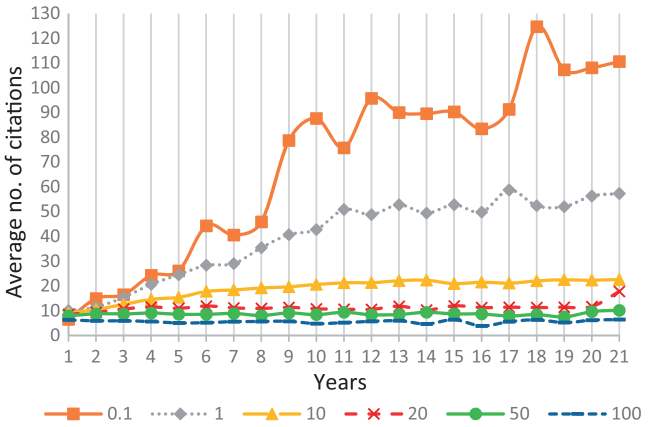

Figure 6. Common references with collaborators over 21 years based on the authors' impact classes.

collaborative with the author or that they work in a completely different research area and the collaborators have little or no theory or methods of value to the author.

The answers to the first two questions, related to Figs 1 and 2, reveal that authors with different levels of impact have different level of citations to, and common references with, their collaborators. This may be related to the structure of science in that social groups in science could be divided into a spectrum ranging from coherent to noncoherent structures based on the level of organization and communication. The two figures suggest that high-impact authors are more engaged in similar lines of research, share more research interest or goals, cite each other's work more, and intellectually and socially associate with each other more compared with the lower-impact authors. Thus, high-impact authors may be relatively organized compared with lower-impact authors.

The answers to the first two questions related to Tables 3 and 4 show that, except for uncited authors, scholars prefer to choose the same references as their collaborators with the same level of impact. They perhaps work on the same set of methods and theories that determine the boundaries of the problems, and they may tend to research related set of problems. The findings also support the invisible colleague hypothesis, which indicates knowledge transfers through informal channels in scientific communities.

In answer to the third research question, authors increasingly cite their collaborators in the long term (Fig. 3) and their total number of collaborations increases linearly over time in the long term (Fig. 4). These findings are consistent with the hypothesis that an author's social environment becomes more and more important as a source of useful information over time, although other explanations are possible for the pattern. Authors as scholars, intentionally or unintentionally, organize and enrich their information environment to get the right information at the right time. Scholars need to be constantly informed about the latest information in their domain and favorite subject areas, because they have information-sensitive tasks and need intellectual and social sustainability. Social and physical communications are essential components of this kind of environment and may provide useful information. In this context, the usefulness of a piece of information is determined by its ability to solve specific classes of problems that could be affected by the common paradigm in which a group of scholars work. Although, collaborators have different degrees of influence, they could be considered as source of information. Authors' social and intellectual ties with collaborators increase gradually over time; however, collaborators as source of information have a core-periphery structure. Authors may be more connected (socially and intellectually) with the core, but the periphery is also important because they may come from other fields, and thus help authors to work on more interdisciplinary problems and conduct research that is more original.

The answers to the third question reveal that the change in the growth rate of common references with, and citations to, collaborators are much greater for high-impact authors than for lower-impact authors. This suggests again that authors in the high-impact classes are more similar to each other compared with other authors and presumably the coauthorship and citation networks of high-impact scientists will show more dense and cohesive connections compared with those of lower-impact authors. Thus, high-impact authors may have different collaboration strategies than do other authors. They rely more on the works of their collaborators over time, and form more coherent research groups in science. These kinds of groups may provide more social support for their research, which may result in higher-impact papers. This is an alternative to the more usual hypothesis that individual projects with more collaborators produce higher quality work.

8. Conclusions

It seems possible that collaboration tends to increase the quality and impact of an article because of the additional knowledge available and it may also increase the impact of an article because of the additional social support for spreading information about it. Nevertheless, the long-term impact of collaboration beyond individual projects needs to be investigated.

An author's number of collaborators gradually increases as they write more papers. Simultaneously, the number of citations to collaborators and the number of common references with them also incrementally increases, but these are limited to some extent and are different in degree from one collaborator to another. Nevertheless, a typical author has no common references with and citations to many of their collaborators. They also have a few common citations with and/or no citations to many of their collaborators, while only a small group of collaborators has a large number of common references with the author and the author cites them much more than other groups of collaborators. Hence, collaboration should not be seen as a uniform scholarly activity but must be viewed as one that can have very different influences on scholars, depending on the nature of the collaborators, and presumably also depending on the nature of the collaboration.

References

Adams, J., Jackson, L. and Marshall, S. (2007) 'Bibliometric analysis of interdisciplinary research'. Report to the Higher Education Funding Council for England.

Adams, J. (2012) 'Collaborations: the Rise of Research Networks', *Nature*, 490/7420: 335–6.

Ahuja, G. (2000) 'Collaboration Networks, Structural Holes, and Innovation: A Longitudinal Study', *Administrative Science Quarterly*, 45/3: 425–55.

Andrade, H. B. *et al.* (2009) 'Dimensions of Scientific Collaboration and its Contribution to the Academic Research Groups' Scientific Quality', *Research Evaluation*, 18/4: 301–11.

Bakshy, E. *et al.* (2012) 'The role of social networks in information diffusion', *Proceedings of the 21st International Conference on World Wide Web*, pp. 519–28. Lyon, France: ACM.

Bigdeli, Z. and Gazni, A. (2012) 'Authors' Sources of Information: A New Dimension in Information Scattering', *Scientometrics*, 92/3: 505–21.

Bigdeli, Z. *et al.* (2012) 'Patterns of Authors' Information Scattering: Towards A Causal Explanation of Information Scattering from a Scholarly Information-seeking Behavior Perspective', *Scientometrics*, 96/9: 1–29.

Bilenko, M., Kamath, B. and Mooney, R. J. (2006, December) 'Adaptive Blocking: Learning to Scale up Record Linkage', *Sixth International Conference of IEEE on Data Mining, 2006, ICDM'06*, pp. 87–96.

Blondel, V. D. *et al.* (2008) 'Fast Unfolding of Communities in Large Networks', *Journal of Statistical Mechanics: Theory and Experiment*, 2008/10: P10008.

Brody, S. (1995) 'Impact Factor as the Best Operational Measure of Medical Journals', *Lancet*, 346/8985: 1300.

Burright, M. A., Hahn, T. B. and Antonisse, M. J. (2005) 'Understanding Information Use in a Multidisciplinary Field: A Local Citation Analysis of Neuroscience Research', *College and Research Libraries*, 66/3: 198–211.

Crane, D. (1969) 'Social Structure in a Group of Scientists: A Test of the" Invisible College" Hypothesis', *American Sociological Review*, 335–52.

Culotta, A. *et al.* (2007) 'Author Disambiguation Using Error-Driven Machine Learning with a Ranking Loss Function', *Sixth International Workshop on Information Integration on the Web (IIWeb-07)*, Vancouver, Canada.

Defazio, D., Lockett, A. and Wright, M. (2009) 'Funding Incentives, Collaborative Dynamics and Scientific Productivity: Evidence from the EU Framework Program', *Research Policy*, 38/2: 293–305.

Didegah, F. and Gazni, A. (2011) 'The Extent of Concentration in Journal Publishing', *Learned Publishing*, 24/4: 303–10.

Didegah, F. and Thelwall, M. (2013) 'Which Factors Help Authors Produce the Highest Impact Research? Collaboration, Journal and Document Properties', *Journal of Informetrics*, 7/4: 861–73.

Fleming, L., Mingo, S. and Chen, D. (2007) 'Collaborative Brokerage, Generative Creativity, and Creative Success', *Administrative Science Quarterly*, 52/3: 443–75.

Franceschet, M. and Costantini, A. (2010) 'The Effect of Scholar Collaboration on Impact and Quality of Academic Papers', *Journal of Informetrics*, 4/4: 540–53.

Fuchs, B. E. *et al.* (2006) 'Behavioral Citation Analysis: Toward Collection Enhancement for Users', *College and Research Libraries*, 67/4: 304–24.

Gazni, A. and Didegah, F. (2011) 'Investigating Different Types of Research Collaboration and Citation Impact: A Case Study of Harvard University's Publications', *Scientometrics*, 87/2: 251–65.

Goldfinch, S., Dale, T. and DeRouen, K. (2003) 'Science from the Periphery: Collaboration, Networks and 'Periphery Effects' in the Citation of New Zealand Crown Research Institutes Articles, 1995-2000', *Scientometrics*, 57/3: 321–37.

Griffith, B. C., Jahn, M. J. and Miller, A. J. (1971) 'Informal Contacts in Science: a Probabilistic Model for Communication Processes', *Science*, 173/3992: 164–66.

Jassawalla, A. R. and Sashittal, H. C. (1998) 'An Examination of Collaboration in High-Technology New Product Development Processes', *Journal of Product Innovation Management*, 15/3: 237–54.

Johnson, B. and Oppenheim, C. (2007) 'How Socially Connected are Citers to Those That They Cite?', *Journal of Documentation*, 63/5: 609–37.

Kanani, P. and McCallum, A. (2007) 'Efficient Strategies for Improving Partitioning-based Author Coreference by Incorporating Web Pages as Graph Nodes', *Proceedings of AAAI 2007 Workshop on Information Integration on the Web*, pp. 38–43. California, USA: AAAI Press.

Kang, I. S. *et al.* (2009) 'On Co-authorship for Author Disambiguation', *Information Processing and Management*, 45/1: 84–97.

Katz, J. S. and Martin, B. R. (1997) 'What is Research Collaboration?', *Research Policy*, 26/1: 1–18.

Krugman, P. R. (1991) *Geography and Trade*. Cambridge, MA: MIT Press.

Kurmis, A. P. and Kurmis, T. P. (2006) 'Exploring the Relationship Between Impact Factor and Manuscript

Rejection Rates in Radiologic Journals', *Academic Radiology*, 13/1: 77–83.

Kuruppu, P. U. and Moore, D. C. (2008) 'Information Use by PhD Students in Agriculture and Biology: A Dissertation Citation Analysis', *Portal: Libraries and the Academy*, 8/4: 387–405.

Lambert, R. (2003) 'Lambert Review of Business-university Collaboration: Final Report', University of Illinois at Urbana-Champaign's Academy for Entrepreneurial Leadership Historical Research Reference in Entrepreneurship.

Lambiotte, R. and Panzarasa, P. (2009) 'Communities, Knowledge Creation, and Information Diffusion', *Journal of Informetrics*, 3/3: 180–90.

Laudel, G. (2002) 'What do We Measure by Co-authorships?', *Research Evaluation*, 11/1: 3–15.

Levin, M. *et al.* (2012) 'Citation-Based Bootstrapping for Large-scale Author Disambiguation', *Journal of the American Society for Information Science and Technology*, 63/5: 1030–47.

Levitt, J. M. and Thelwall, M. (2010) 'Does the Higher Citation of Collaborative Research Differ from Region to Region? A Case Study of Economics', *Scientometrics*, 85/1: 171–83.

Liu, Y., Rousseau, R. and Guns, R. (2013) 'A Layered Framework to Study Collaboration as a Form of Knowledge Sharing and Diffusion', *Journal of Informetrics*, 7/3: 651–64.

Maglaughlin, K. L. and Sonnenwald, D. H. (2005) 'Factors that Impact Interdisciplinary Natural Science Research Collaboration in Academia', *Proceedings of the Conference of the International Society for Scientometrics and Informetrics*, pp. 499–508. Stockholm, Sweden.

Mattsson, P., Sundberg, C. J. and Laget, P. (2011) 'Is Correspondence Reflected in the Author Position? A Bibliometric Study of the Relation Between Corresponding Author and Byline Position', *Scientometrics*, 87/1: 99–105.

McRae-Spencer, D. M. and Shadbolt, N. R. (2006) 'Also by the Same Author: AKTiveAuthor, a Citation Graph Approach to Name Disambiguation', *Proceedings of the 6th ACM/IEEE-CS joint conference on Digital libraries*, pp. 53–4. Chapel Hill, NC, USA: ACM.

Morris, S. A. and Van der Veer Martens, B. (2008) 'Mapping Research Specialties', *Annual Review of Information Science and Technology*, 42/1: 213–95.

Moya-Anegón, F. *et al.* (2013) 'The Research Guarantors of Scientific Papers and the Output Counting: A Promising New Approach', *Scientometrics*, 97/2: 421–34.

Ohniwa, R. L. *et al.* (2004) 'Perspective Factor: A Novel Indicator for the Assessment of Journal Quality', *Research Evaluation*, 13/3: 175–80.

Porac, J. F. *et al.* (2004) 'Human Capital Heterogeneity, Collaborative Relationships, and Publication Patterns in a Multidisciplinary Scientific Alliance: A Comparative Case Study of Two Scientific Teams', *Research Policy*, 33/4: 661–78.

Price, D. D. S. (1986) *Little Science, Big Science and Beyond*. New York: Columbia University Press.

Reagans, R. and McEvily, B. (2003) 'Network Structure and Knowledge Transfer: The Effects of Cohesion and Range', *Administrative Science Quarterly*, 48/2: 240–67.

Reuther, P. and Walter, B. (2006) 'Survey on Test Collections and Techniques for Personal Name Matching', *International Journal of Metadata, Semantics and Ontologies*, 1/2: 89–99.

Riesenberg, D. and Lundberg, G. D. (1990) 'The Order of Authorship: Who's on First?', *JAMA*, 264/14: 1857.

Rinia, E. J. *et al.* (2001) 'Citation Delay in Interdisciplinary Knowledge Exchange', *Scientometrics*, 51/1: 293–309.

——. (2002) 'Measuring Knowledge Transfer Between Fields of Science', *Scientometrics*, 54/3: 347–62.

Schrage, M. (1995) *No More Teams!: Mastering the Dynamics of Creative Collaboration*. New York, NY: Currency Doubleday.

Singh, J. (2005) 'Collaborative Networks as Determinants of Knowledge Diffusion Patterns', *Management Science*, 51/5: 756–70.

Smalheiser, N. R. and Torvik, V. I. (2009) 'Author Name Disambiguation', *Annual Review of Information Science and Technology*, 43/1: 1–43.

Smith, L. C. (1981) 'Citation Analysis', *Library Trends*, 30/1: 83–106.

Soler, J. M. (2007) 'A Rational Indicator of Scientific Creativity', *Journal of Informetrics*, 1/2: 123–30.

Song, Y. *et al.* (2007) 'Efficient Topic-based Unsupervised Name Disambiguation', *Proceedings of the 7th ACM/IEEE-CS Joint Conference on Digital Libraries 7*, pp. 342–51. Vancouver, Canada.

Sonnenwald, D. H. (2007) 'Scientific Collaboration', *Annual Review of Information Science and Technology*, 41/1: 643–81.

Stokes, T. D. and Hartley, J. A. (1989) 'Coauthorship, Social Structure and Influence Within Specialties', *Social Studies of Science*, 19/1: 101–25.

Thagard, P. (1997) 'Collaborative Knowledge', *Noûs*, 31/2: 242–61.

Tijssen, R. J. and Van Leeuwen, T. N. (2006) 'Measuring Impacts of Academic Science on Industrial Research: A Citation-based Approach', *Scientometrics*, 66/1: 55–69.

Torvik, V. I. *et al.* (2005) 'A Probabilistic Similarity Metric for Medline Records: A Model for Author Name Disambiguation', *Journal of the American Society for Information Science and Technology*, 56/2: 140–58.

Wallace, M. L., Gingras, Y. and Duhon, R. (2009) 'A New Approach for Detecting Scientific Specialties from Raw Cocitation Networks', *Journal of the American Society for Information Science and Technology*, 60/2: 240–6.

Wallace, M. L., Larivière, V. and Gingras, Y. (2012) 'A Small World of Citations? The Influence of Collaboration Networks on Citation Practices', *PloS One*, 7/3: e33339.

Wilson, P. (1996) 'Interdisciplinary Research and Information Overload', *Library Trends*, 45/2: 192–203.

Wilson, T. D. (2006) 'On User Studies and Information Needs', *Journal of documentation*, 62/6: 658–70.

Wuchty, S., Jones, B. F. and Uzzi, B. (2007) 'The Increasing Dominance of Teams in Production of Knowledge', *Science*, 316/5827: 1036–9.

Yan, E. and Ding, Y. (2012) 'Scholarly Network Similarities: How Bibliographic Coupling Networks, Citation Networks, Cocitation Networks, Topical Networks, Coauthorship Networks, and Coword Networks Relate to Each Other', *Journal of the American Society for Information Science and Technology*, 63/7: 1313–26.

Yan, E. and Sugimoto, C. R. (2011) 'Institutional Interactions: Exploring Social, Cognitive, and Geographic Relationships Between Institutions as Demonstrated Through Citation Networks', *Journal of the American Society for Information Science and Technology*, 62/8: 1498–514.

Yan, E. *et al.* (2013) 'A Bird's-eye View of Scientific Trading: Dependency Relations Among Fields of Science', *Journal of Informetrics*, 7/2: 249–64.

Zhang, L. *et al.* (2010) 'Subject Clustering Analysis Based on ISI Category Classification', *Journal of Informetrics*, 4/2: 185–93.