The small world of citations: How close are citing authors to those they cite?

Matthew L. Wallace[1], Vincent Larivière[2] and Yves Gingras[3]

[1] *matt.l.wallace@gmail.com*
Centre interuniversitaire de recherche sur la science et la technologie (CIRST), Université du Québec à Montréal, Case Postale 8888, Succursale Centre-Ville, Montréal, Québec, H3C 3P8 (Canada) and
Science Policy Division, S&T Branch, Environment Canada, 200 Sacré-Coeur Blvd, 11th Floor, Gatineau, Quebec, K1A 0H3 (Canada)

[2] *lariviere.vincent@uqam.ca*
Observatoire des sciences et des technologies (OST), Centre interuniversitaire de recherche sur la science et la technologie (CIRST), Université du Québec à Montréal, Case Postale 8888, Succursale Centre-Ville, Montréal, Québec, H3C 3P8 (Canada) and
Cyberinfrastructure for Network Science Center, School of Library and Information Science, Indiana University, 10th Street & Jordan Avenue, Wells Library, Bloomington, Indiana, 47405 (United States)

[3] *gingras.yves@uqam.ca*
Observatoire des sciences et des technologies (OST), Centre interuniversitaire de recherche sur la science et la technologie (CIRST), Université du Québec à Montréal, Case Postale 8888, Succursale Centre-Ville, Montréal, Québec, H3C 3P8 (Canada)

Abstract
This analysis examines the proximity of authors to those they cite using degrees of separation in a co-author network, expanding on the notion of self-citations. When rigorously computed using all cited and citing authors, the proportion of direct self-citations are relatively constant in time across various specialties in the natural sciences (10% of citations) and the social sciences (20%). Citations to nearby authors of the co-author network, however, vary widely among fields and time periods. Authors in specialties such as astrophysics and astronomy, for instance, have very few citations outside their network of collaborators. We discuss, in social and mathematical terms, the degree to which this closeness is determined by the size and topology of the co-author network (especially as it is affected by recent increases in co-authorship) and by the referencing practices of different disciplines. These results have implications for the long-standing debate over biases common to most types of citation analysis, and especially for understanding social structures and citation practices across scientific disciplines over the past 50 years. In addition, our findings have important practical implications for the availability of 'arm's length' expert reviewers of grants applications and manuscripts.

Introduction
Great strides have been made in understanding the function and nature of referencing within articles. References can provide symbolic capital to the cited author(s), can be selected in terms of a strategy of *persuasion*, or to demonstrate allegiance (or emphasize differences) with respect to a subset of a given specialty (Gilbert, 1972). Though studies have shown that persuasion is dominant, it is difficult to separate and distinguish the motivations of referencing (Brooks 1986). References can also be divided into classes based on their cognitive relationship with the cited work: basic, subsidiary, additional and perfunctory (Chubin & Moitra, 1975). In fact, Gilbert (1972) remarks that the latter subset, which includes the majority of self-citations and 'social' citations, is "puzzlingly large".
A common argument against such use of citations in research evaluation is the presence of self-citations, as they could be increased by the authors themselves (MacRoberts & MacRoberts, 1989). Similar concerns are related to the formation of 'citation cartels' or to cronyism, which purportedly serve to modify a journal's impact factor, an institution's citation count or an author's h-index (Franck, 1999; Phelan, 1999). However, there is a lack of large-scale empirical data on if and where such biases are in fact occurring.

Using a very large dataset (more than 2,6M papers and 50M references) over a 50-year period, this paper combines and expands on methods for analyzing co-author networks and methods for measuring self-citations. It poses the all-important question of whether the social network of researchers has an impact on the selection of references found in a given article. In contrast to White, Wellman and Nazer (2003), who used survey data to characterize a small social network of researchers, we use co-authorship as an indicator of their social proximity. More specifically, we analyze the references of each article in terms of four levels of closeness, loosely based on the concept of Erdös numbers (see Methods section below). In order to distinguish between a variety of citation practices within the natural and medical sciences (NMS) and social sciences and humanities (SSH), eight specialities, based on the NSF classification of journals, were chosen. The following section provides a broad review of the literature on co-author networks, social proximity and self-citations. It is followed by a detailed description of the methods and database used, the results obtained and a discussion of these results.

Literature review

Co-author networks and social proximity

Previous attempts to examine citations in terms of social closeness are sparse. White *et al.* (2003) combined, for a small group of researchers, bibliometrics with survey data to see whether citations were influenced by the social structure of the group. Introducing the notion of 'inter-citation' as a measure of citations between members of a given group, they aimed to correlate citations with social, socio-cognitive and intellectual ties. While the first type is beyond the reach of the present study (and—in general—difficult to accurately measure for the purposes of social network analysis, as discussed in Newman, 2001), these socio-cognitive ties, as defined by co-authorship, will be the focus of the present paper. Intellectual ties, as White *et al.* and many others have shown, are essentially a given when analyzing citation patterns. White et al.'s conclusions (2003), based on 16 individuals, are nuanced: there is some correlation, as one might expect, between collaboration and citation patterns but, overall, there is no strong or reliable link between social ties and citations, nor is there any attempt to cite one another in order to boost reputations. Using a very small dataset, Johnson and Oppenheim (2007) also find a correlation between social ties and citations. Finally, Rowlands (1999) was somewhat successful in complementing a co-citation network with information about whether intellectually close authors knew each other.

While not social networks *per se*, collaboration networks can be seen as a proxy measure for understanding social ties, where large publication databases can provide a wealth of quantitative information (Newman, 2001). More importantly, they are useful for understanding various aspects of the social structure of science. Imbedded in the topology of co-author networks are the positions of its members within the hierarchy of a given subfield: primary investigators would be central hubs, and co-authorship links between groups can reflect residues of individual career paths (Velden *et al*., 2010). Co-authorship networks are not only scale-free, but also value-laden; links can imply different types of connections between authors. James Moody's work (2004) provided a great deal of insight into many such questions in his study of collaboration networks in sociology. The collaboration pattern reveals boundaries between specialties, and suggests likely models of how consensus within the network might emerge on a local scale. Similarly, Newman's extensive examination of large-scale collaboration networks has effectively provided the foundation for a quantitative understanding of co-authorship networks (Newman, 2001, 2004). Like Moody, the focus is on macroscopic properties of the co-author network: the size of the largest component, the distance between authors, etc. This allows one to clearly distinguish between the co-author

network topology in various disciplines. The clustering coefficient, on the other hand, reveals more about the local topology (through density of triangles in the network).

Self Citations

Self-citations were calculated very early in the history of scientometrics and the SCI. Garfield and Sher's 1963 paper calculated that 8% of citations were first-author self-citations, while 20% of citations received by a journal were self-citations. Most of these studies follow the typology of Lawani (1982), which distinguishes between diachronous self-citations (received) and synchronous self-citations (made). Recent studies usually use the method of Snyder and Bonzi (1998), which considers as a self-citation any intersection between the authors of the citing authors of a paper and the authors of the cited paper. Given that Thomson Reuters' Web of Science only indexes the names of co-authors of cited papers that are also *source* items, this definition of self-citation can only be used for references made to source items.

Studies of various disciplines have found rates of self-citations among references varying between 10% and 36%, with strong variations between specialties (Tagliacozzo, 1977, Lawani, 1982, MacRoberts & MacRoberts, 1990), and much lower percentages in SSH such as sociology and economics (Bonzi and Snyder, 1990). Diachronous studies of citations found higher values of between 26% and 37% (Asknes, 2003, Costas *et al*., 2010) at the author or document levels. It has also been found (Glänzel et al., 2006) that 1) self-citations are generally younger and have a shorter half-life than foreign citations, 2) Self-citations stabilize in a period of 3-4 years after publication, 3) there is a clear relationship between the number of citations received and the number of self-citations received and 4) the percentage of self-citations only slightly increases with the number of co-authors.

Finally, the literature reveals that self-citations can actually prove very useful as a heuristic for understanding scientific fields. When paired with information such as keywords or co-author network, self-citation networks can be used to detect emerging fields, or to understand an author's network or career (Hellsten *et al*., 2007). Self-citation also plays an important role in providing credibility and assisting in the promotion of new results and should not necessarily viewed as a 'perverse' aspect of scholarly publication (Hyland, 2003; Glänzel *et al.,* 2006).

Methods

There are two main methodological challenges to answering our question. First, at the expense of detailed qualitative information about social networks, we need to conduct a large-scale analysis to measure the social proximity of referenced authors across scientific disciplines. Second, unlike most bibliometric work, the analysis needs to be centred on the individual authors or papers, in order to gain insight into their referencing practices and individual social networks. The methods used here provide an expansion of our understanding of the sociology of science based on bibliometric methods. However, the present study focuses only on contemporary networks and is centred on the behaviour of individuals within a co-author network.

The data for this analysis comes from Thomson Scientific's Web of Science, which includes the Science Citation Index Expanded (SCIE), Social Science Citation Index (SSCI), and Arts and Humanities Citation Index (AHCI) for the 1945–2008 period. Data is presented for 8 specialities (5 from the NMS, 3 from the SSH) based on the NSF field classification[22]: astronomy and astrophysics, atmospheric science and meteorology, biochemistry and molecular biology, economics, history, neurology and neurosurgery, organic chemistry and sociology. Only research articles, notes and reviews are included in the set.

[22] http://www.nsf.gov/statistics/seind06/c5/c5s3.htm#sb1

In order to investigate the social properties of a given scientific specialty, we form a set of references R_k cited by a set of papers S_k published in a given year k within a given specialty. We restrict this set of references (and their source items S) to those whose source can be identified within the database (i.e. source items), and which were published within the previous 10 years. We generate a list of authors a_k having contributed to each article $s_k \in S_k$, yielding a total set of authors A_k for the specialty as a whole.

Similarly, we generate a second set of authors a'_k (and A'_k for the entire specialty) who collaborated within 2 years[23] (semi-arbitrarily defined) of the publication year k with authors in a_k (restricted to the specialty in question in order to limit false positives due to the presence of homonyms). Thus, A'_k constitutes the unweighted and undirected co-author network. Finally we generate a third group of authors a''_k who collaborated with a'_k during the same time period. It should be noted that a''_k excludes all authors contained in a'_k, so in general, for networks which are relatively sparse, or which containing few numbers of co-authors, $n(A''_k) < n(A'_k)$ (while the opposite is true for cases when collaboration rates are high).

For each article source article $s \in S$, we examine its set of references and classify them in the following way:

A) If any of the authors of the referenced paper is contained in a_k, then this is a **direct self-reference**;

B) If any of the authors of the referenced paper is contained in a'_k, then this is a **reference to a level-1 co-author**;

C) If any of the authors of the referenced paper is contained in a''_k, then this is a **reference to a level-2 co-author**;

D) If none of the authors in the referenced paper are contained in a_k, a'_k, or a''_k, then this is called **distant reference**.

These categories are defined as mutually exclusive: if a referenced paper can be placed in more than one category, then it is assigned the one closest to a self-citation. Many will recognize this is essentially the beginnings of the Erdös number or six degrees of separation game, applied to each author individually and as each of his referenced authors as the 'object' of the game (not always Paul Erdös or Kevin Bacon...). From a sociological perspective, it is not necessary to continue past the second 'level' (Erdös number of 2), since we can consider that there is no longer any social connection between the authors within a given specialty. In addition, given the number of authors and references being considered, the data mining procedure is both expensive in CPU time and memory usage. Figure 1 provides a visual representation of this algorithm.

[23] Because of this ±2 year interval, the data presented is for the 1947-2006 period, though data is collected for the 1945-2008 period.

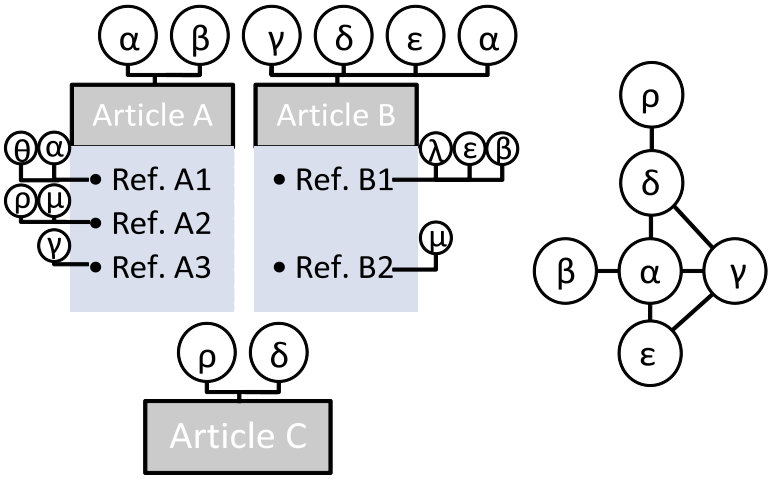

Figure 1: An illustrative representation of the algorithm. Left: a set of three articles and 5 references therein. Right: The corresponding co-author network. Article A, for example is written by two authors (α and β) and contains three references (whose authors are also denoted by Greek letters). Based on our classification scheme of social proximity, references A1 and B1 are *direct references*, A2 is a *reference to a level-2 co-author* (since α collaborated with δ who collaborated with ρ), A3 is a *reference to a level-1 co-author* and B2 is a *distant reference*.

Finally, it should be noted that while the source items and authors are restricted to a given specialty, the items they cite are not. One would expect that the specialty in question covers the majority of 'peers' cited, but such a limit would, while defining a 'closed' system, introduce a social artefact, particularly for more interdisciplinary specialties such as biochemistry (see Figure 2C). However, in such cases, we have checked that the results are similar, whether or not we restrict the specialty of the reference items.

One of the main advantages of our method for examining referencing patterns is the ability to conduct the analysis at the level of each author or paper. It is thus useful to think of each author making referencing *choices* based in part on other authors that are in proximity to him or her. Specifically, given the number of references and authors associated with a given paper, we can consider how many 'close' references (as per our definitions) they pick, compared to those expected randomly. We can define this quantity as the propensity P_d for a given level of proximity d, approximated as the ratio of the number of articles found empirically to the expected number of articles to be found given a random selection of references. The latter is nothing but a binomial distribution, so for a single article in a given year k, the propensity of having a reference to a level-1 or level-2 co-author can be written as:

$$P_{l1,k} = \frac{n(A_k')}{\sum_{i=1}^{n(r_k)} a_i a_k'} n_{l1} \quad ; \quad P_{l2,k} = \frac{n(A_k'')}{\sum_{i=1}^{n(r_k)} a_i a_k''} n_{l2} \qquad (1)$$

where n_{l1}, n_{l2} are the number of cases empirically identified at each level, $n(a_i)$ the number of authors of the reference i, $n(r_k)$ the number of 'remaining' references of the given paper. Like our data presented in Figure 3, the propensity is computed in sequence, in order of proximity, with the 'matched' references removed at each step. In other words, the level-2 propensity, for instance, is not 'skewed' by the number of direct or level-1 references already found for the given paper. Thus, the numerator is determined by empirical 'matches' and the size of the entire network, while the denominator reflects size of the author's network and that of the cited authors' networks.

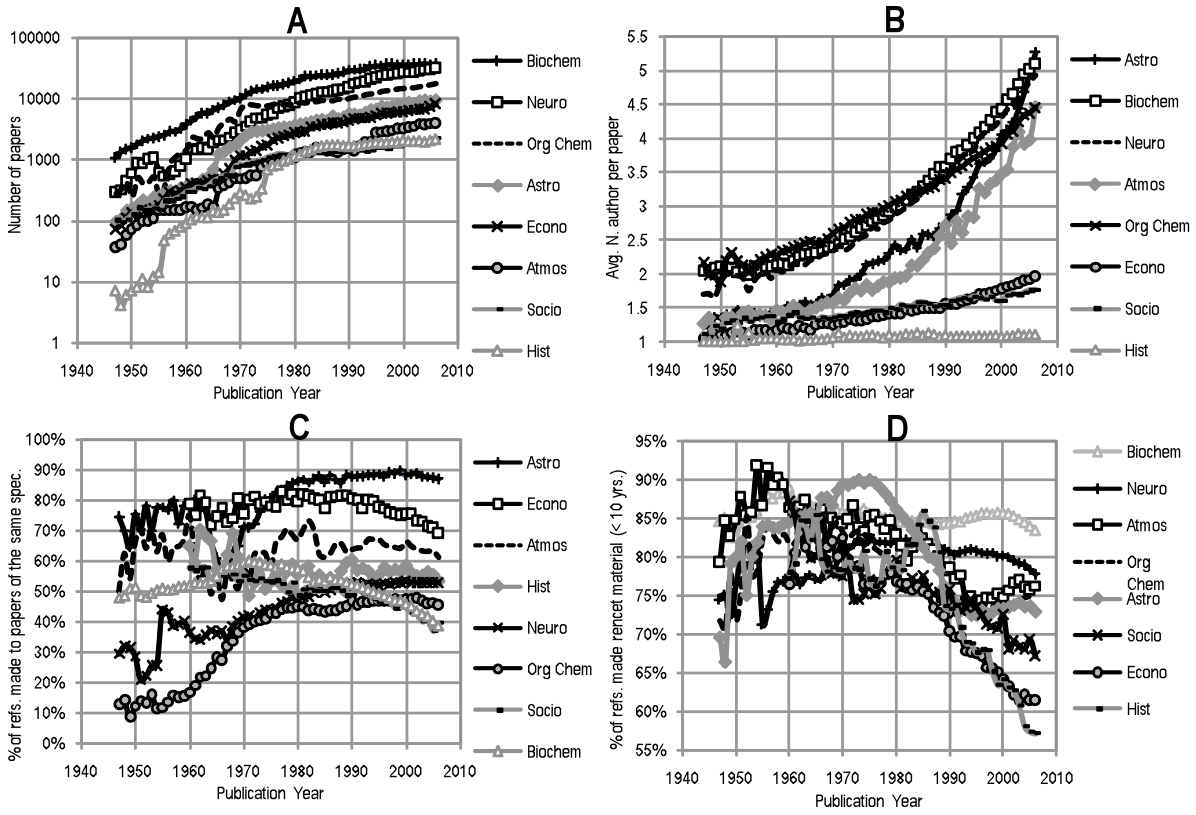

Figure 2. For each of the chosen specialties, A) number of papers, B) average number of authors per paper, C) percentage of identified references within the same specialty, and D) percentage of identified references that defined as 'recent' (less than 10 years older than the source item).

Results

For a number of specialties, we first compute the percentage of references within the given specialty (e.g., economics to economics), essentially expanding on measurements of journal self-citations. We also calculate a few basic macroscopic variables (number of papers, number of authors, etc.) characterizing the field. This allows us to select a representative sample of specialties (with different rates of 'intra-citations', different sizes, levels of co-authorship, etc.) as shown in Figure 2.

Figure 3 shows the distribution of citations across several specialties in the NMS (A-E) and the SSH (F-H). As one might expect, the proximity of references in each of the disciplines varies a great deal. Within the natural sciences, one immediately notices a major difference in the social proximity of references between, on the one hand, astrophysics/astronomy and atmospheric science and meteorology, and the rest of the specialties on the other hand. Aside from organic chemistry, all specialties show a clear decrease in the percentage of references made to 'distant' papers (to authors to which they have no connection). Furthermore, while it is clear that the size of the specialties (Figure 2A), the number of co-authors per paper (Figure 2B), and the proportion of 'intra-specialty' references (Figure 2C) have a clear impact on the closeness of references (as one might expect), none of these macroscopic quantities can singlehandedly explain the trends observed in Figure 3. In addition, there is no major correlation between the tendency to cite recent literature (Figure 2D) and the proportion of that literature that is socially 'proximate'.

Direct self-citation, however, is relatively constant both across fields and over time, hovering around 20% in NMS specialties and 10% in the SSH. The difference between the NMS and SSH is substantial, and dwarf the differences among SSH specialties shown in Figure 3. We find that there is no such thing as 'group' self-citations, defined as citations made to previous

co-authors, within the three SSH fields studied. This primarily due to the fact that co-authorship is less frequent in these disciplines and that, as a consequence, researchers have less co-authors in their social network to choose from, a clear limitation in the way we define our social network. For this reason, the rest of our article mostly focuses on the NMS.

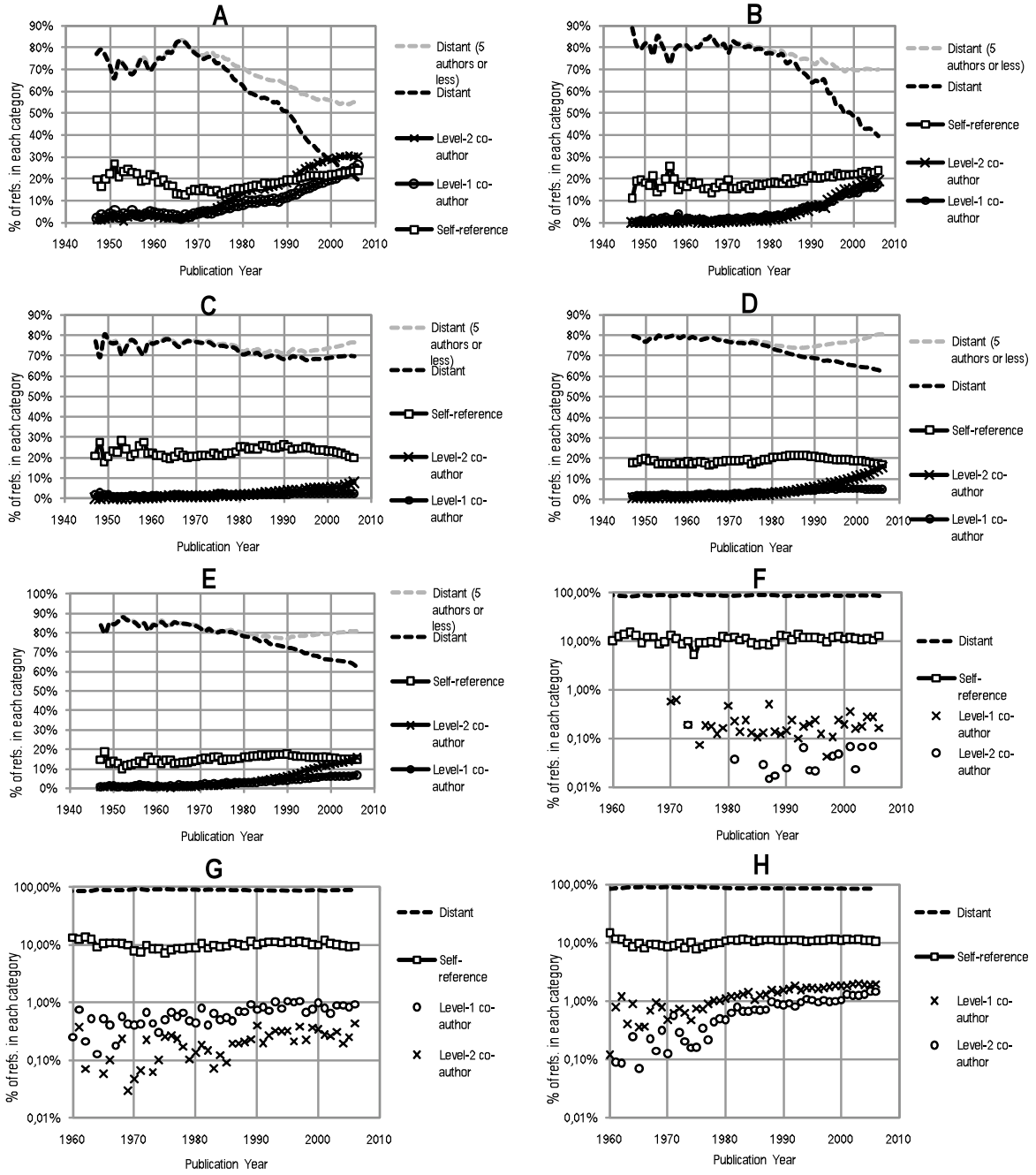

Figure 3. The distribution of references contained based on proximity for five natural science and three social science specialties: A) astrophysics and astronomy, B) atmospheric science and meteorology, C) organic chemistry, D) biochemistry and molecular biology, E) neurology and neurosurgery, F) history, G) sociology and H) economics. The last three are shown on a logarithmic scale for clarity. For the NMS, we compute the same distribution based on a subset of source articles (and their references) which contain only 5 authors or less (dashed gray lines).

For NMS disciplines, we also show the corresponding distribution of references when we limit the set *S* for each year to papers with 5 co-authors or less (gray dashed lines in Figure 3).

While arbitrary, this immediately gives us a sense of the extent to which disciplines such as astrophysics and astronomy have high degrees of social proximity in their references due to the presence of papers with high levels of co-authorship. Furthermore, it is more likely that authors of papers with 5 authors or less actually *know* each other. For clarity, we omit from Figure 3 the number of self-references, references to level-1 co-authors and to level 2 co-authors when this restriction is imposed. Interestingly, the increase in 'distant' references observed is at the expense of references to level-1 and level-2 co-authors, but not to direct self-references.

Reducing the number of co-authors to 5 or less is not sufficient to understand to what degree the number of other authors in proximity to a given author influences their choices. This begs the question of how the number of level-1 and level-2 co-authors is distributed within each of the specialties. Figure 4 shows these distributions for two periods: 1960-1969 and 2000-2006. Two main observations can be made. First, variations in distributions of co-authors do not correlate highly with variations differences in the number of 'close' citations (Figures 3A-E). Second, the relatively even distribution of level-2 co-authors means that, within a given network, there will be wide variations in how many of these more distant co-authors are 'available' to a given author.

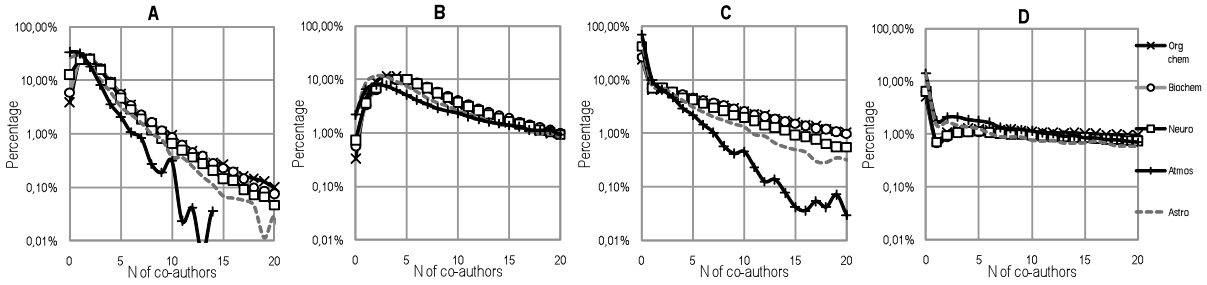

Figure 4. Distribution of the number of A) level-1 co-authors ($a_k^{'}$) and C) level-2 co-authors ($a_k^{''}$) during the 1960-1969 period; B) level-1 co-authors ($a_k^{'}$) and D) level-2 co-authors ($a_k^{''}$) during the 2000-2006 period.

We have confirmed that distributions of co-authors are likely too 'coarse' and cannot account for all increases in the proximity of citations, by bringing the distributions of authors per paper in each of the specialties closer together for all articles published between 2004 and 2006. This is done by randomly removing source papers (up to around 15% of the network in order to maintain its 'shape') until the distributions of authors per paper are almost identical[24] and using only the first author of references. Similarly, we can randomly remove papers in a given specialty such that each author in a given interval of time has only 1 paper. These procedures have the effect of diluting the network (i.e., reducing the amount of clusters) (Newman, 2001, 2004). Once again, we see no major effect on the closeness of citations.

[24] In practice, it is difficult to make the different sets of source papers have exactly the same distribution of authors per paper. Our objective is to reduce the effect of skewed distributions while ensuring that the 'reduced' network retains sociological meaning.

Table 1. Propensity for citation to level-1 (top) and level-2 (bottom) co-authors.

P_{l1}	*Astro*	*Atmos*	*Biochem*	*Econo*	*Hist*	*Neuro*	*Org chem*	*Socio*
1956-65	20	6	76	19		18	30	6
1966-75	50	27	139	26	5	56	56	22
1976-85	59	41	168	50	24	93	58	18
1986-95	65	54	155	74	20	96	46	35
P_{l2}	*Astro*	*Atmos*	*Biochem*	*Econo*	*Hist*	*Neuro*	*Org chem*	*Socio*
1956-65	4.0	1.6	12.5	0.4		2.2	6.4	0.6
1966-75	7.9	8.6	20.9	2.2	0.4	7.9	10.9	2.0
1976-85	8.6	7.2	21.6	8.8	0.2	15.6	11.4	1.8
1986-95	8.7	7.9	15.9	11.2	1.1	13.5	6.9	3.2

Finally, Table 1 shows the propensity (computed individually for each source paper in the 10-year period, then averaged) for references to level-1 and level-2 co-authors for three time periods, as described above in the Methods section. In general, there is very little propensity to reference level-2 co-authors. Data for direct self-citations (not shown) are an order of magnitude higher than for level-1 co-authors, as one might expect. Furthermore, there appears to be an overall plateau in propensity to cite 'close' authors as of 1975 or so.

Discussion

Self-citations and 'group' self-citations

This remarkable stability in the level of direct self-referencing—across specialties and time—distinguishes this practice from that of referencing those who are 'close'. This suggests that there are cross-cutting norms regarding this practice in science. It must be noted that this does not imply a degree of conformity *within* the specialization (comparisons of the distribution of self-citations would reveal the degree to which actors adhere to this norm). However, the stability of the average is important in understanding that this practice is not 'social' nor random (i.e., it does not depend on the number of co-authors), but is a widespread and stable practice in all disciplines. For this reason and due to the increasing importance of research groups as a dominant unit for understanding scientific work, it is important to expand the notion of self-citations to 'group' self-citations which reflect social proximity. Many of these would be captured by a rigorous (all co-authors to all co-authors) definition of self-references as used here, but not all. We do not dispute that studying direct self-citations can be useful as a bibliometric heuristic (on the scale of a few authors) or that it is a crucial part of the publication process for scientists, only that it cannot reveal anything, on the aggregate level, about the overall referencing practices or structure of a given scientific specialty.

Combinatorial effects and the social structure of scientific specialties

When expanding the notion of self-citations to a given author's co-authors (and co-authors' co-authors), the effect of having a large number of collaborators per paper is amplified. Our findings clearly show that recent increases in the proximity of citing and cited authors are, in part, due to an increase in the size of collaborations. This is the case in astrophysics and astronomy, for instance. Co-authorship practices in fields such as astrophysics or particle physics often reflect the use of certain instruments or of a willingness to acknowledge the contributions of a wider range of individuals in the division of labour, beyond the writing of the article itself. In this sense, there is a sociological basis to this combinatorial effect.

Our results also clearly show that the combinatorial effect cannot alone account for the proximity of citing and cited authors. Indeed, from a social network perspective, the co-author network is defined by more than the distribution of edges per node. In other words, it is not

just about how large collaborations are, but also of what type of collaborations occur and where. We have also found that the distribution of clustering coefficients (Watts & Strogatz, 1988) is very similar in the co-author networks of five NSM scientific specialties in recent years. This essentially measures the concentration of triangles within the network or to what extent collaborators of a given author also collaborate with each other. Therefore, other measures must be able to account for the local structure of the networks. Along the lines of Moody (2004), we view self-citation and 'group' citation as a means to reinforce local social networks, which has particular importance for the intellectual and social development of scientific specialties.

The issue of 'compact' vs. 'fragmented' fields can only be partially explored through co-author networks, and intellectual structures are generally better identified through co-citation analysis. We must be cautious in extrapolating our results to the intellectual structure of these specialties: while two research groups may be entirely disconnected in terms of social proximity, they may be working on identical topics. Since our 'experiment' is at the local level, we can only really speak of 'micro-fragmentation' at the level of small clusters. Once again, we know that large collaborations in astrophysics (e.g., around telescopes) and atmospheric sciences (e.g., around general circulation models) naturally create large hubs that dominate the network. Whether or not these hubs are linked and part of the same component of the network is largely irrelevant to the present study, since we operate at the local level. These topological effects are generally correlated with a concentration of references to the same specialty, to certain groups of journals or to certain central research groups. Recent studies of astrophysics, for instance, have confirmed the trends observed here of an increased reliance on a small number of journals (Abt, 2009).

How do authors choose which peers to cite?

We have shown that the topology of the network does not account for all differences between fields in the proximity of citing and cited authors. We wanted to know if, despite the *availability* of papers written by 'distant' peers, authors in specialties such as atmospheric science and meteorology choose to rely predominantly on the scholarship of those who are 'closer'. The propensity (Equation 1 and Table 1, above) quantifies the factor by which authors cite level-1 and level-2 co-authors more often than what we could *expect* given the local structure of their specialty. This allowed us to account for the fact that, in fields where (at the local level) there are many co-authors (and many co-authors cited), there is a higher probability of obtaining 'close' references.

If a relatively large field (e.g., biochemistry and molecular biology, or economics) contains many groups working on *largely independent* topics, then the propensity tends to be high. It also indicates that high levels of 'close' references in astrophysics and astronomy, or in atmospheric science and meteorology, are largely due to the structure of the specialties, not the choices made by citers. Thus comparing practices in different specialties (SSH included) with very different co-author network topologies shows that level-1 citations are far from random, which likely reflects the specialization of researchers and the cumulative nature of research. Interestingly, the only two specialties which, recently, tend to cite fewer 'close' authors are organic chemistry and, to a lesser degree, biochemistry. This confirms the validity of the trend observed earlier in Figure 3 and might indicate either that different types of referencing practices exist within organic chemistry (e.g., there are less perfunctory references) or that the field is less intellectually fragmented and authors search out information from further afield.

Conclusion

Our new method combining self-citations with a network-based sociology of science allow us to broadly characterize the social closeness of citations in the sciences. Most importantly, the vast quantities of data allow for an unambiguous comparison of the proximity of citations among specialties. It is found that self-referencing is a relatively stable practice across different specialties, both in the SSH and NMS. The same cannot be said for citations to collaborators or collaborators of collaborators. So the answer to the question posed in the title would be "it depends on the social context and citation practices of the specialty in question". As a general rule, however, the purported 'small-worldness' only applies to 'close' individuals (i.e., recent co-authors) and to only a few scientific specialties.

There is no single key to understanding why authors of a given specialty may cite authors with whom they, or their co-authors, have previously published. Our results highlight the importance of a few main factors that determine to what degree this takes place: the level of intra-specialty referencing (to what degree does scholarly work build on a closed set of journals), the level of collaboration (particularly for very large collaborations) and the propensity of individual authors to cite work from within their social network (given a local network of a certain shape). This last factor is a particularly important indicator of actors' citation practices within a given scientific specialty.

Our findings could be complemented by a closer investigation of the context of these references, the nature of the co-authorships (whether the authors are colleagues, students, in the same research group, etc.), the specific motivations of proximate citations and the perception thereof within different fields. In addition, our analysis does not examine the degree of homogeneity in the 'closeness' of references (or the propensity) within each field.

In the context of a broader understanding of trends in the structure and practices of the various NMS and SSH specialty areas, our analysis points globally to the presence of more close-knit research groups in many fields, some increased bias towards 'social' perfunctory references, and an increased fragmentation of research topics and groups. Recent work regarding the decline of uncitedness (Wallace *et al*., 2009) and strong evidence that scholarship is becoming less and less 'concentrated' (Larivière *et al*., 2009) points to the fact that scholarship is not 'narrowing' within science in general, although our data shows a correlation between fields' high levels of 'close' references and high levels of intra-specialty citations (Figures 2 and 3).

Proximate referencing is generally regarded as a perversion of the citation process, and seen as evidence that a field is too inward-looking or controlled by a small number of authors. Our analysis suggests that this is not necessarily the case. Furthermore, co-authorship itself can have many meanings, not only in terms of division of labour, but also as a means of establishing a hierarchy within a field. The formation of large groups who feed off each other's ideas and periodically collaborate does not necessarily imply citation cartels or nepotism. However, it is true that the socio-cognitive 'compactness' of fields such as astrophysics and astronomy, or meteorology and the atmospheric sciences, might pose certain problems. For instance, it can be more difficult to locate 'unbiased', arm's length reviewers of papers and grants, or may make it more challenging for unknown authors to get recognized within a given area of study.

References

Abt, H.A. (2009). Do astronomical journals have extensive self-referencing? *Publications of the Astronomical Society of the Pacific*, 121, 73-75.

Aksnes, D.W. (2003). A macro study of self-citation. *Scientometrics*, 56, 235–246.

Bonzi, S. & Snyder, H. (1990). Patterns of self citation across fields of inquiry. *Proceedings of the 53rd Annual Meeting of the American Society for Information Science*, 27, 204–207.

Brooks, T. A. (1986), Evidence of complex citer motivations. *Journal of the American Society for Information Science*, 37, 34-36.

Chubin, D.E. & Moitra, S. (1975). Content Analysis of References: Adjunct or Alternative to Citation Couting? *Social Studies of Science*, 5, 423-441.
Costas, R. van Leeuwen , Th.N. & Bordons, M. (2010) Self-citations at the meso and individual levels: effects of different calculation methods. *Scientometrics*, 82, 517–537.
Franck, G. (1999). Science communication—A vanity fair? *Science*, 286 (5437), 53-55.
Frandsen, T.F. (2007). Journal self-citations - Analysing the JIF mechanism. *Journal of Informetrics*, 1, 47-58.
Garfield, E. & Sher, I.H. (1963). New factors in the evaluation of scientific literature through citation indexing. *American Documentation*, 18, 195–201.
Gilbert, G.N. (1977). Referencing as persuasion. *Social Studies of Science*, 7, 113-22.
Glänzel, W., Debackere, K., Thijs, B. & Schubert, A. (2006) A concise review on the role of author self-citations in information science, bibliometrics and science policy. *Scientometrics*, 67, 263-277.
Hellsten, I., Lambiotte, R., Scharnhorst, A. & Ausloos, M. (2007). Self-citations, co-authorships and keywords: A new approach to scientists' field mobility? *Scientometrics*, 72, 469-486.
Hyland, K. (2003). Self-citation and self-reference: Credibility and promotion in academic publication. *Journal of the American Society for Information Science and Technology* 54, 251-259.
Johnson, B., Oppenheim, C. (2007). How socially connected are citers to those that they cite? *Journal of Documentation,* 63, 609-637.
Larivière, V., Archambault, É. & 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, 288-296.
Larivière, V., Gingras, Y. & Archambault, É (2009). The decline in the concentration of citations, 1900-2007. *Journal of the American Society for Information Science and Technology*, 60, 858-862.
Lawani, S.M. (1982). On the Heterogeneity and Classification of Author Self-Citations. *Journal of the American Society for Information Science*, 33, 281-284.
MacRoberts, M. H. & MacRoberts, B. R. (1989). Problems of citation analysis: A critical review. *Journal of the American Society for Information Science*, 40, 342-349.
Moody, J. (2004). The Structure of a Social Science Collaboration Network: Disciplinary cohesion from 1963 to 1999. *American Sociological Review*, 69, 213-238.
Newman, M.E.J. (2001). Scientific collaboration networks. I. Network construction and fundamental results. *Physical Review E*, 64, 016131.
Newman, M.E.J. (2004). Coauthorship networks and patterns of scientific collaboration. *Proceedings of the National Academy of Sciences*, 101, 5202-5205
Phelan, T. J. (1999). A compendium of issues for citation analysis. *Scientometrics*, 45, 117-136.
Rowlands, I. (1999). Patterns of author cocitation in information policy: evidence of social, collaborative and cognitive structure. *Scientometrics*, 44, 533-546.
Snyder, H. & Bonzi, S. (1998). Patterns of self-citation across disciplines (1980-1989). *Journal of Information Science*, 24, 431–435.
Tagliacozzo, R. (1977). Self-citation in scientific literature. *Journal of Documentation*, 33, 251-265.
Velden T., Haque, A. & Lagoze, C. (2010). A new approach to analyzing patterns of collaboration in co-authorship networks: mesoscopic analysis and interpretation. *Scientometrics,* 85, 219-242.
Wallace, M.L., Larivière, V. & Gingras, Y. (2009). Modeling a Century of Citation Distributions. *Journal of Informetrics*, 3, 296-303.
Watts, D.J. & Strogatz, S.H. (1998). Collective dynamics of 'small-world' networks. *Nature*, 393, 440-442.
White, H.D., Wellman, B., Nazer, N. (2004). Does citation reflect social structure? Longitudinal evidence from the "Globenet" interdisciplinary research group. *Journal of the American Society for Information Science and Technology*, 55, 111–126.