Researchers' publication patterns and their use for author disambiguation

Vincent Larivière and Benoit Macaluso
École de bibliothéconomie et des sciences de l'information, Université de Montréal, C.P. 6128, Succ. Centre-Ville, Montréal, Québec, H3C 3J7, Canada and Observatoire des sciences et des technologies (OST), Centre interuniversitaire de recherche sur la science et la technologie (CIRST), Université du Québec à Montréal, Case Postale 8888, Succursale Centre-Ville, Montréal, Québec, H3C 3P8, Canada
vincent.lariviere@umontreal.ca; macaluso.benoit@uqam.ca

Abstract

Over the recent years, we are witnessing an increase of the need for advanced bibliometric indicators on individual researchers and research groups, for which author disambiguation is needed. Using the complete population of university professors and researchers in the Canadian province of Québec (N=13,479), of their papers as well as the papers authored by their homonyms, this paper provides evidence of regularities in researchers' publication patterns. It shows how these patterns can be used to automatically assign papers to individual and remove papers authored by their homonyms. Two types of patterns were found: 1) at the individual researchers' level and 2) at the level of disciplines. On the whole, these patterns allow the construction of an algorithm that provides assignation information on at least one paper for 11,105 (82.4%) out of all 13,479 researchers—with a very low percentage of false positives (3.2%).

Introduction

Since the creation of the Science Citation Index in the 1960s—and the subsequent online availability of Thomson's various citation indexes for the sciences, social sciences and the humanities through the Web of Science (WoS)—most large-scale bibliometric analyses have mainly been performed using the address (institutions, countries, etc.) journal, paper or discipline field. Analyses made with the author field are much scarce, and typically used small sample of researchers[1]. There is, thus, an important part of the bibliometric puzzle that was missing: the individual researcher, to which we can attribute socio-demographic characteristics (gender, age, degree, etc.). Until the last few years, the issues related to the assignation of papers to individuals had not been discussed extensively in the bibliometric community (Enserink, 2009). However, the advent of h-indexes (Hirsch, 2005) and its numerous variants (Egghe, 2006; Schreiber, 2008; Zhang, 2009) aimed at evaluating individual researchers, as well as the need for more advanced bibliometric data compilation methods for measuring the research output of research groups whose names do not appear on papers (e.g. interuniversity groups, departments, etc.) or for measuring the effect of funding on researchers' output and impact (Campbell *et al.*, 2010), has increased the interest for bibliometric data on individuals and, hence, for author disambiguation.

The main challenge of author-level analyses is the existence of homonyms (or the inexistence of a researcher unique identifier), which makes the assignation of papers to distinct individuals quite difficult. Two general types of problems can be found at the level of authors (Smalheiser and Torvik, 2009). First and foremost, two or several individual can share the same name (homonyms). Secondly, one researcher can

[1] When research was performed at the level of authors,. The recent collection of Scientometrics papers dealing with individual researchers published by Academia Kiado (Braun, 2006) illustrates this trend: the study with the highest number of researchers included has less than 200. Similarly, notable studies in sociology of science by the Coles (1973), Merton (1973) and Zuckerman (1979) analyzed small datasets.

sign papers in more than one manner (with or without initial(s), maiden name, etc.). These difficulties are exacerbated by two characteristics of the Web of Science (WoS). First, prior to 2006, only the surname and initial(s) of authors' first name(s) were indexed, for a maximum of three initials. Hence, researchers sharing the same surname and initial(s)—for example, John Smith and Jane Smith—were grouped under the same distinct string (Smith-J). Although the complete given name of authors is now indexed in the WoS, it only does so for journals providing this information in the author section of their papers[2], in addition to the fact that it obviously does not solve the problem for papers published before 2006. Similarly, prior to 2008, no link was made in the database between an author and its institutional address. Although this was not a problem for sole authored papers—which only represent a slight fraction of papers published—it was more problematic for co-authored papers. More specifically, for a paper authored by three researchers and on which three institutional addresses are signed, it is impossible to know the exact institution of affiliation of each author, as several combinations are possible. Hence, the search for 'Smith-J' among papers on which McGill University appears will, for example, retrieve papers from John Smith and Jane Smith, but also from Joseph Smith who, albeit not from McGill University, has collaborated with an author from McGill (homonymy of collaborators). There is, however, still a dearth of information on the extent of the homographic problem in the scientific community. Apart from Aksnes (2008) and Lewison (1996), who respectively compiled data on the extent of homonyms among Norwegian researchers and on the frequency of author's initial(s)—but did not test directly their effect on the compilation of bibliometric data on individual researchers—there is very little information on the extent to which researchers share the same name and its effect on the compilation of bibliometric data at the level of individual researchers.

This papers aims at contributing to this literature by presenting regularities found in papers manually assigned to the entire population of university professors and researchers (N=13,479) in the province of Québec (Canada)[3] as well as all the papers that were authored by their homonyms. It first reviews some of the relevant literature on the topic, and then presents a series of regularities found in researchers' publication patterns and how these can be used to automatically assign papers to individual researchers. Two types of patterns are presented: 1) individual researchers' past publication behavior and how it determines subsequent behaviour and 2) the relationship between researchers' departmental affiliation and the disciplines in which they publish. These patterns are then used, in a reverse engineering manner, to automatically assign papers to individuals. Results in terms of both false positives and false negative are presented and discussed in the conclusion.

Previous studies' on the assignation of individual authors' publications

Over the last few years, several studies have provided algorithms for the disambiguation of individual researchers. However, most studies – with the notable exception of Reijnhoudt et al. (2013) and Levin *et al*. (2012) have been performed using relatively small datasets (Gurney, Horlings and van den Besselaar, 2012; Wang et al., 2012) and, quite often, without actually having clean data on the papers authored by homonyms and papers authored by the "real" researcher. Jensen *et al.* (2008) attempted to compile publication and citation files for 6,900 CNRS researchers using the Web of Science. Instead of removing, in each researchers' publication file, the papers written by homonyms, they evaluated the probability that a given researchers has homonyms, and, if this probability was high, they completely removed the researcher from the sample. More precisely, they first measured, by comparing the surname and initials of each

[2] Physics journals for instance, often having very long author lists, do not index the complete given name(s) of authors.

[3] See for example Gingras *et al.* (2008) and Larivière *et al.* (2010) for the some results based on this population.

researcher (VLEMINCKX-S) with some of its variants (VLEMINCKX-SG, VLEMINCKX-SP, etc.), the probability that the researcher has homonyms. If the researcher had too many variants, it was removed. Their second criterion was related to the number of papers published: if a researcher had too many papers, it was considered as an indication that more than one scientist was behind the records. Hence, researchers had to publish between 0.4 and 6 papers per year to be considered in the sample. Their third criterion was that the first paper of each researcher to have been published when the researcher was between 21 and 30 years old. Their resulting database contained 3,659 researchers (53% of the original sample).

This method has at least two major shortcomings. First, the fact that a name (e.g. VLEMINCKX-S) is unique does not imply that it represents only one distinct researcher. In this particular case, it could be the surname and initial combination for Serge Vleminckx, Sylvain Vleminckx, Sophie Vleminckx, etc. Second, it removes from the sample highly active researchers (who published more than 6 papers per year), which obviously distorts their results. This method is similar to that of Boyack and Klavans (2008), who used researchers with uncommon surnames to reconstitute individual researchers' publication and patenting activities. Using the combination of the name of the author/inventor and the research institution[4] signed on the paper, they calculated the odds that the paper belonged to the given author.

Another method is that of Han *et al.* (2005) who, using K-means clustering algorithms and Naïve Bayes probability models, managed to categorize 70% of the papers authored by the very common 'strings' Anderson-J and Smith-J into distinct clusters. The variables they used were the names of co-authors, the name of journals and the title of the papers. The assumption behind this algorithm is that researchers generally publish papers on the same topics, in the same journals and with the same co-authors. A similar method was also used by Torvik *et al* (2005) using Medline. Similarly, Wooding *et al.* (2006) used co-authors for removing homonyms from a sample of 29 principal investigators funded by the Arthritis Research Campaign. For each author, they first found a core of papers which, without a doubt, belonged to the right researcher. Using this 'core' subset of papers in the specialty of arthritis, they created, for each researcher, a list of co-authors which were used to gather papers in areas other than arthritis. A novel aspect of this study is that several rounds of co-author inclusion were performed, increasing between each round the number of co-authors in the core. After three rounds of the algorithm, 99% of the authors' papers were assigned—which could be considered as the recall of papers—with only 3% of false positives (97% precision). This method is very similar to that used by Kang *et al.* (2009), and has been expanded by Reijnhoudt et al., (2013), to include additional heuristics (such as email address and reprint author, among others). Cota *et al.* (2010) also used similar heuristics (co-author, title of paper and publication venue) and manage to disambiguate authors of about 4,500 papers of the DBLP and BDBComp collections.

Aswani, Bontcheva and Cunningham (2006) used, in addition standard bibliographic information (abstract, initials of the authors, titles and co-authors), automatic web-mining for grouping papers written by the same author. The Web-mining algorithm searches for the full names of the authors, tries to find their own publications' page, etc. Their results show that web-mining improves the clustering of papers into distinct authors, but the small sample used in the study makes the results less convincing. On the whole, most of these studies indeed manage to 1) automatically disambiguate authors or to 2) automatically assign papers to authors, although most of them do so with very small datasets, and often without a thorough analysis of false-negative, false-positives and various error rates. Finally, Levin et al. (2012) developed a citation-based bootstrapping algorithm – with an emphasis on self-citations – to disambiguate 54 million WoS author-paper

[4] The bibliometric part of their paper used the Scopus database, which, contrary to Thomson Reuters' databases, links names of authors with institutional addresses since 1996.

combinations. They show that, when combined with emails, author names, and language, self-citations was the best bootstrapping element. They then manually disambiguated 200 authors – which, in this context, is not a large sample -- to assess the precision and recall of the algorithm, and found values of 0.832 and 0.788, respectively.

Methods

Contrary to most existing studies on the topic, this study uses, as a starting point, a list of distinct university based researchers (N=13,479), including their department and university (Larivière *et al.*, 2010). The database on university researchers' papers and of those authored by homonyms was thus obtained by matching the surname and initials of these researchers contained in the list to the surname and initials of authors of Quebec's scientific articles indexed in the Web of Science[5]. This first match resulted in a database of 125,656 distinct articles and 347,421 author-article combinations. Each article attributed to each researcher was then manually validated in order to remove the papers authored by homonyms. This manual validation is generally made by searching the title of each of the papers on Google to find their electronic versions on which, generally, the complete names of the authors are written. This often helps to decide if the papers belong to the researcher. Another method is to search the name of the researcher on Google to find his/her website to get an indication of is publications' list or CV. After a few papers, one generally understands the publication pattern of the researcher and correctly attributes his/her papers. This essential but time-consuming step reduced the number of distinct papers by 51% to 62,026 distinct articles and by 70% to 103,376 author-article combinations. Analysis of this unique dataset, including the characteristics of both assigned and rejected papers, sheds light on the extent of homonyms in Quebec's scientific community.

To assess the reliability and reproducibility of the manual validation of university researchers and professors' publication files, tests with different individual 'attributors' were performed for a sample of 1,380 researchers (roughly 10% of the researchers). It showed that for most publication files, the two coders manually assigned exactly the same papers. More specifically, 1,269 files (92%) researchers had exactly the same papers assigned. A difference of one paper was found in 72 cases (5.2%), 2 papers in 15 cases (1.1%), 3 papers in 9 cases (0,7%) and 4 papers in 3 (0,2%) cases. The remaining 12 files had a maximum difference of 12 papers each. In terms of author-article combinations, the error rates are even lower. Out of the 12,248 author-article links obtained the first time, 12,124 (or 99%) remained unchanged the second time. Manual validation is thus quite reliable and reproducible.

In order to find patterns in researchers' publication output, this study uses, for each of Quebec's university researcher and professor, a dataset of all the WoS-Indexed papers with authors that matched their authors' name (for example, Smith-J) amongst all papers with at least one Canadian address for the 2000-2007 period. These papers were manually categorized as belonging to the right researcher or as belonging to a homograph, which allows—contrary to most studies presented in the preceding section—to test how these patterns could be used to discriminate false positives from papers that rightly belong to a researcher. The difference between researchers' papers that were manually assigned and those rejected allows the testing of the algorithm.

[5] It is worth noting that, while the WoS does not have a build-in algorithm for author disambiguation, Scopus has one, for which the results are available directly on the Scopus web interface. However, there is not information available on the method used.

In order to help the search for patterns, each journal indexed in the WoS was assigned a discipline and a specialty according to the classification scheme used by U.S. National Science Foundation (NSF) in its Science and Engineering Indicators series (Appendix 1)[6]. The main advantage of this classification scheme over that provided by Thomson Reuters is that 1) it has a two level classification (discipline and specialty), which allows the use of two different levels of aggregation and, 2) it categorizes each journal into only one discipline and specialty, which prevents double counts of papers when they are assigned to more than one discipline. Similarly, a discipline was assigned to each of the researchers' department (Appendix 2). These disciplines were assigned based on the 2000 revision of the U.S. Classification of Instructional Programs (CIP) developed by the U.S. Department of Education's National Center for Education Statistics (NCES)[7]. This dataset serves as the backbone for finding the relationships between the disciplinary affiliation of university researchers and the discipline of their publications.

Regularities in researchers' publication patterns

A first interesting piece of information found in this dataset is the percentage of papers of each researcher that were retained after manual validation. More specifically, these *cleaned* publication files made it possible to estimate the extent of homonyms problems for all Quebec university researchers for whom at least one article was automatically matched (N= 11,223) using the name of the researcher within papers having at least one Quebec institutional address[8]. With an automatic matching of researchers' names, compared to a cleaned publication file (Figure 1):

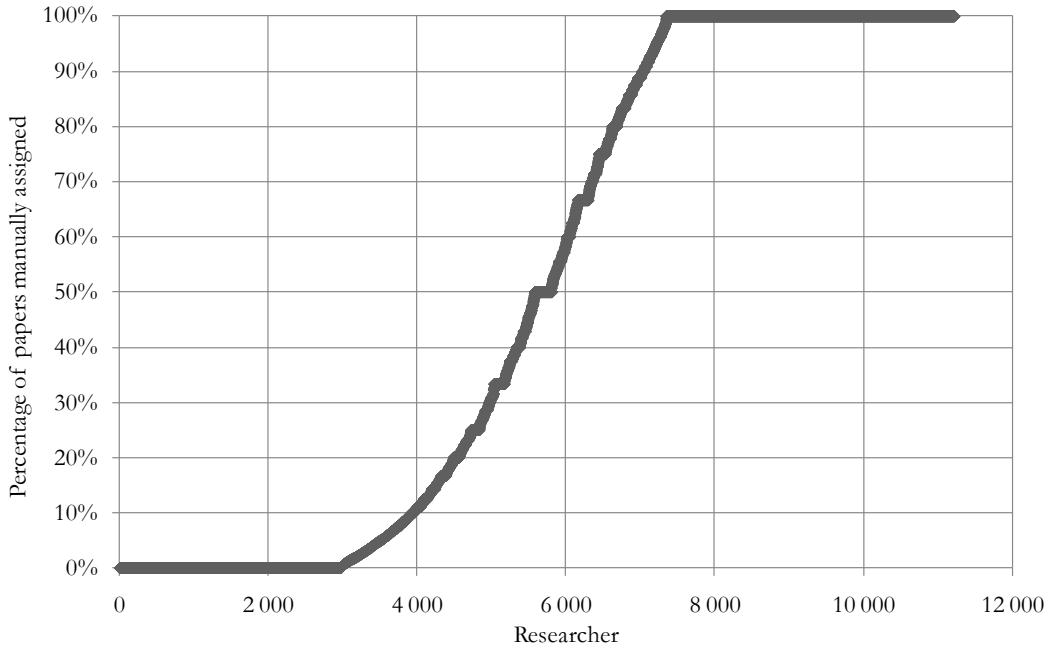
- The papers matched for 2,972 researchers (26.5%) were all rejected which, in turn, meant that they had not actually published any papers (all papers were written by homonyms);
- Between 0.1 and 25% of the papers matched were assigned for 1,862 researchers (16.6%);
- Between 25.1 and 50% of the papers matched were assigned for 975 researchers (8.7%);
- Between 50.1 and 75% of the papers matched were assigned for 722 researchers (6.4%);
- Between 75.1 and 99.9% of the papers matched were assigned for 818 researchers (7.3%);
- The papers matched of 3,874 researchers (34.5%) were all conserved after manual validation (i.e., they had no homonyms within the subset of Quebec papers).

[6] More details on the classification scheme can be found at: http://www.nsf.gov/statistics/seind06/c5/c5s3.htm#sb1

[7] For more details on the CIP, see: http://nces.ed.gov/pubs2002/cip2000/.

[8] Thus, 2,256 of Quebec's researchers did not publish any paper during that period nor had any of their homonyms.

Figure 1. Percentage of papers assigned after manual validation, by researcher



Crude matching—without removing papers authored by homonyms—is thus valid for slightly more than a third of the researchers. On the other hand, the scientific production of the remaining two-thirds was significantly overestimated. Since it is impossible to know *a priori* which researchers will be overestimated and which ones will not, the validation of each paper from each researcher is, theoretically, needed. As mentioned previously, papers of these publication files were all manually validated (assigned or rejected) and serve, in a reverse engineering manner, as a test bed for finding patterns in the publications of researchers.

In a manner similar to that of Wooding *et al.* (2006) for arthritis research, papers were then analyzed in order to find characteristics which could help isolating a core of papers for each researcher—i.e. a subset of all of each researcher' paper that we are sure are not those of homonyms. This was more complex in the context of this paper, as core papers had to be found for researchers that could be active in any field of science and not only in arthritis. After several rounds of empirical analysis, the combination of three variables optimized the ratio between the number of papers found and the percentage of false positives. Figure 2 and 3 present the two sets of criteria with which a core set of papers could be found for university-based researchers. Figure 2 present the first matching criteria: the complete name of researchers matched with the complete name of authors—including the complete given name (available in the Web of Science since 2006)—and the name of the researcher's university matched with the name of the university on the paper.

Figure 2. First matching criteria for creation the core of papers

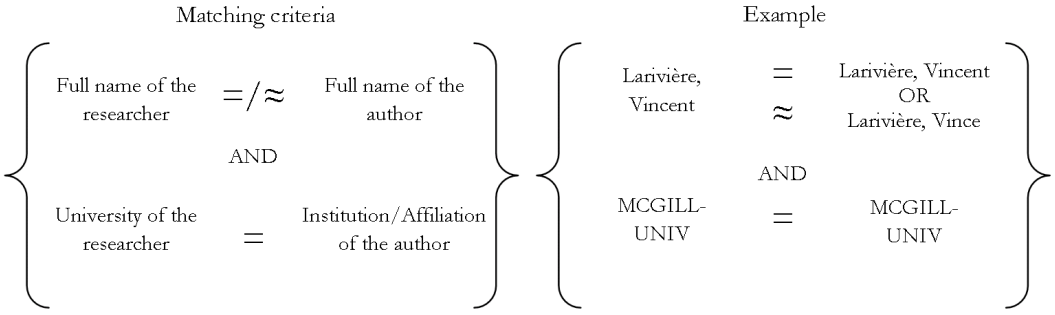
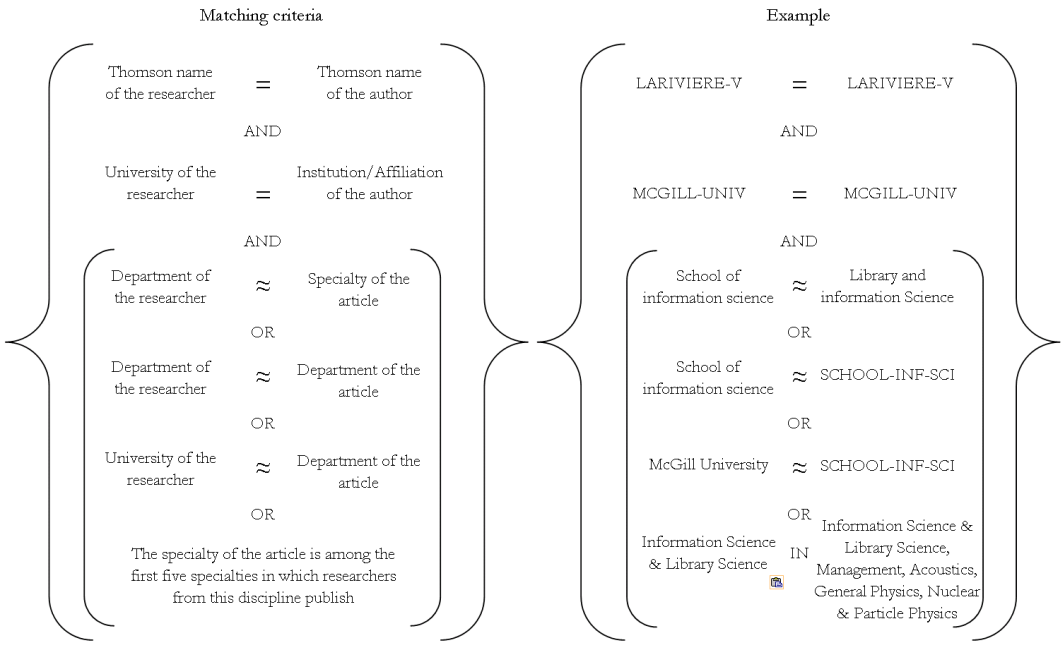


Figure 3 presents the second matching criteria. Firstly, the name of the author of the paper had to be written exactly in the same manner as the name of the researcher in the list. Secondly, the institution appearing on the paper (or its affiliation, e.g. Royal Victoria Hospital is affiliated to McGill University) had to be the same as the institution appearing on the list and, thirdly, the discipline of the journal in which the paper is published, the department or the institution of the authors had to be similar[9] to the department of the researcher as it appeared on the list of university professors and researchers or the discipline of the paper had to be among the five disciplines in which researchers from this department published most.

Figure 3. Second matching criteria for creation the core of papers



Following Boyack and Klavans (2008), an analysis of rare surnames was also performed, which were defined as surnames only belonging to one individual in the list of university researchers. Hence, all papers

[9] The similarity threshold (MinSimilarity) was set at 0.25 in Microsoft SQL Server SQL Server Integration Services (SSIS). More details on the system can be found at: http://technet.microsoft.com/en-US/library/ms345128(v=SQL.90).aspx

authored by researchers having a rare name, and on which their institution of affiliation appears, were included in core papers. As shown in Table 1, these three criteria allow the creation of a core of papers for more than 75% of the individual researchers for which at least one paper has been manually assigned (8,081), matches 56.4% of their distinct papers and 47.5% of the author-paper combinations, e.g. LARIVIERE-V and paper 'X'. At each level of analysis, the number of false positives is rather low; and is especially low at the level of author-paper combinations (less than 1%).

Table 1. Results of the matching of core papers at the levels of university researchers, articles and author-paper combinations

Unit of analysis	Manual validation (N)	Automatic assignation		False positives	
		N	%	N	%
Researchers	8 081	6 117	75,7%	344	4,3%
Articles	62 629	35 353	56,4%	772	1,2%
Author-paper combinations	97 850	46 472	47,5%	809	0,8%

Another set of regularities was found in individual researchers' publication patterns. The idea behind this search for patterns for individual researchers was to be able, using subset of papers in the core, to find other papers that belonged to the researchers but that did not exhibit the characteristics found in Figures 2 and 3. To do so each researcher's publication record was divided into two distinct time periods: 2000 to 2003 and 2004 to 2007.

Using the characteristics of the papers published by each given researcher during the first time period, it was then tried to automatically assign to the same researcher the papers published during the second time period. Two indicators were quite successful in doing so: the use of the same words in the title, author keywords and abstract fields of upcoming publications (Figure 4) and the citing of the same references (Figure 5) of papers for which the Thomson name [e.g. LARIVIERE-V] and the institution [MCGILL-UNIV] also matched. Figure 4 presents the percentage of rightly and wrongly attributed papers, as a function of the keyword index. The keyword index is a simple indicator compiled for each 2004-2007 paper matched to a researcher, based on the keywords of the papers assigned to the researcher for the period 2000-2003. Its calculation is as follows:

$$Ki_{pr} = \left(\frac{N_{kpm}}{N_{kp}} \times \frac{1}{\sqrt{N_{kt}}}\right) \times 100 \qquad \text{Eq. (1)}$$

Where N_{kpm} is the number of keywords of a 2004-2007 paper that match the keywords used in the 2000-2003 papers of a researcher, N_{kp} is the total number of keywords of the 2004-2007 paper and N_{kt} is the total number of keywords used in all the 2000-2003 papers assigned to the researcher. The square root of N_{kt} was used instead of N_{kt} alone in order to obtain an overall number of keywords (denominator) that is

not too high—especially for very productive researchers. The result is multiplied by 100 in order to be closer to an integer.

Figure 4 shows that when the keyword index is at two, about 90% of the papers rightly belong to the researcher and that, to the opposite, slightly greater than 10% are false positives. When the keyword index is greater than 2 (3 or more), the percentage of rightly assigned papers rises above 95%, and stays at this level until 7, where about 100% of the papers are assigned to the right researcher. These numbers mean that it is possible to rightly assign papers to a researcher using the regularities found in the title words, keywords and words of the abstract.

Figure 4. Percentage of rightly assigned and wrongly assigned papers, as a function of the keywords previously used by a university researcher, 2000-2003 and 2004-2007

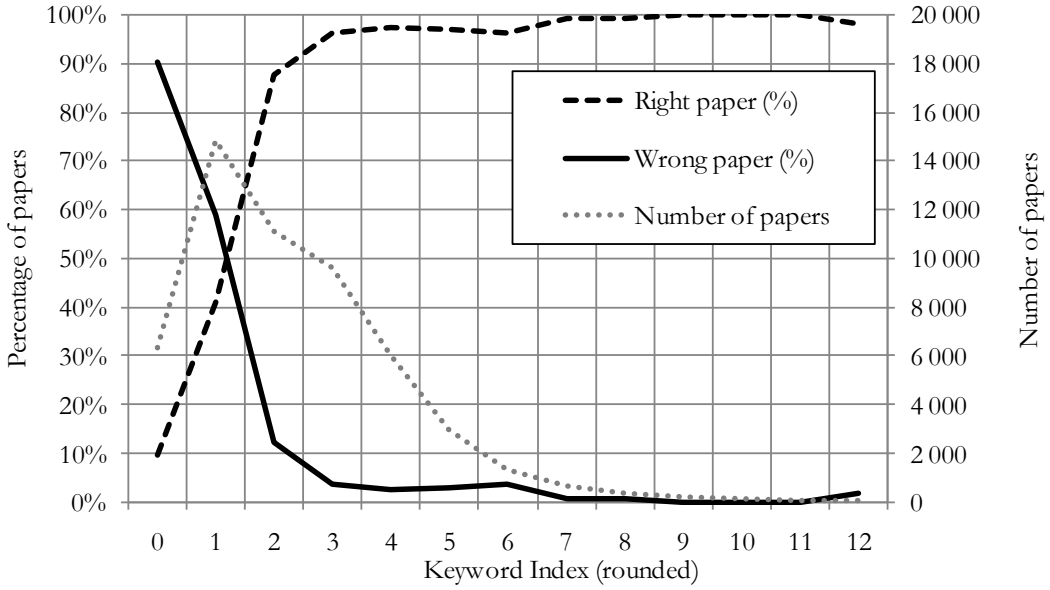


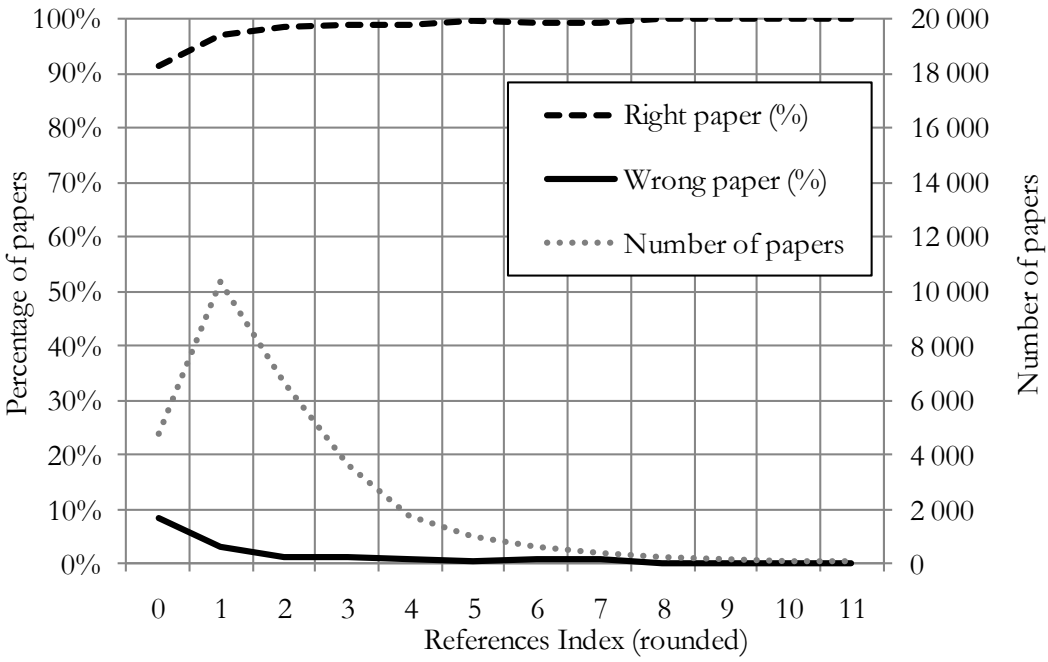
Figure 5 presents the references index for 2004-2007 papers, based on papers published between 2000 and 2007. The references index is very similar to the keyword index previously presented; it is based on the pool of references made previously (2000-2003) by the researcher. Its calculation is as follows:

$$Ri_{pr} = \left(\frac{N_{rpm}}{N_{rp}} \times \frac{1}{\sqrt{N_{rt}}}\right) \times 100 \qquad \text{Eq. (2)}$$

Where N_{rpm} is the number of references of a 2004-2007 paper that match the references used in the 2000-2003 papers of a researcher, N_{rp} is the total number of references of the 2004-2007 paper and N_{rt} is the total number of references used in all the 2000-2003 papers assigned to the researcher. Again, The square root of N_{rt} was used instead of N_{rt} alone in order to have an overall number of cited references (denominator) that is not too high. The result is also multiplied by 100 in order to be closer to an integer.

Figure 5 shows that as soon as a signal is obtained, i.e. that at least one of the referenced work of the 2004-2007 paper was previously made in the 2000-2003 dataset, more than 90% of the papers rightly belong to the researcher. When the references index increases to 1 or above, the quasi-totality the papers rightly belong to the researcher.

Figure 5. Percentage of rightly assigned and wrongly assigned papers, as a function of the references previously made by a university researcher, 2000-2003 and 2004-2007



Using the keywords and references found in the papers assigned in core (set at 2 or more for the keyword index and at >0 for the references index), we then assigned 10,892 additional papers were assigned, with only 236 papers being false positives (2.2% of the added papers), for an overall error rate at the level of papers of 2.2% and of 1.7% at the level of author-paper combinations (Table 2). Since this matching of papers can only be made for researchers for which a certain number of core papers were matched, the number of researchers stays the same, but slightly more researchers have at least one paper wrongly assigned (6.7%).

Table 2. Results of the matching of core papers and papers with the same keywords or cited references, at the levels of university researchers, articles and author-paper combinations

Unit of analysis	Manual validation (N)	Automatic assignation		False positives	
		N	%	N	%
Researchers	8 081	6 117	75,7%	407	6,7%
Articles	62 629	46 245	73,8%	1 008	2,2%
Author-paper combinations	97 850	64 765	66,2%	1 078	1,7%

Another round of automatic matching of papers was also performed with the same references and keywords (set at the same thresholds), but using only the Thomson name [e.g. LARIVIERE-V] and the province [QC], but not the institution [MCGILL-UNIV]. Using this method, 3,645 additional papers were retrieved, of which 674 were false positives (Table 3). Although this percentage seems quite high, the overall proportion of false positives at the level of articles remains quite low (3.2%) and is even lower for author-paper combinations (2.3%).

Table 3. Results of the matching of core papers and papers with the same keywords or cited references, without the 'same institution' criteria, at the levels of university researchers, articles and author-paper combinations

Unit of analysis	Manual validation (N)	Automatic assignation		False positives	
		N	%	N	%
Researchers	8 081	6 117	75,7%	576	9,4%
Articles	62 629	49 890	79,7%	1 577	3,2%
Author-paper combinations	97 850	72 918	74,5%	1 682	2,3%

In order to increase the number of researchers for which a certain number of core papers could be found, the relationship between the discipline of researchers and the specialty of papers was analyzed. An increase in the number of researchers for which core papers could be found is important because core papers are the starting point of the automatic assignation of several other papers. For each of the 5,615 existing combinations of disciplines of publications (Appendix 1) and of disciplines of departments (Appendix 2), a matrix of the percentage of papers from each discipline of publication that rightly belonged to researchers from each department was calculated. Unsurprisingly, it was found that papers published in the main specialty in which researchers from a given specialty publish were more likely to belong to the right researchers. For example, 100% of the 186 papers published in geography journals that matched the names of authors of geography departments belonged to the right researcher. The same is true for several other obvious department-specialty relationships, such as university researchers from chemical engineering departments publishing in chemical engineering journals (99% of the 1,017 papers rightly assigned), but also for less obvious relationships such as researchers in civil engineering publishing in Earth & planetary science journals (95% of the 316 papers rightly assigned).

On the other hand, all the 333 papers published in biochemistry & molecular biology journals that matched authors from the disciplines of anthropology, archaeology & sociology belonged to the wrong researcher. The same is also true for the 202 papers published in organic chemistry that matched authors from business departments. Given that no university-affiliated researcher from this domain has ever published in journals of this specialty during the period studied, there are low chances that researcher of the same domain will do so.

Figure 6 presents the matrix of the percentage of assigned papers, for each combination of discipline of departments (x-axis) and specialty of publication (y-axis). Darker zones are combinations of specialties of publications and disciplines of departments where a larger proportion of papers were accepted during manual validation; lighter zones are combinations where a majority of papers were rejected during manual validation. This figure illustrate that there is a majority of discipline of department/specialty of publication combinations where the quasi totality of papers were authored by homonyms (light zones), and a few darker zones where a large proportion of papers were accepted. Unsurprisingly, zones where most of the papers were assigned are generally cases where the discipline of the department is related with the discipline of the journal—for example, researchers from departments of information science & library science publishing in journals of library & information sciences. The presentation of this landscape clearly shows that there are some combinations where the majority of papers were assigned during manual validation and others where only a minority of papers was assigned during the process. We can, thus, focus on these light zones to automatically exclude papers from a given department published in given specialty, and on dark zones to automatically include papers from other department/specialty combinations.

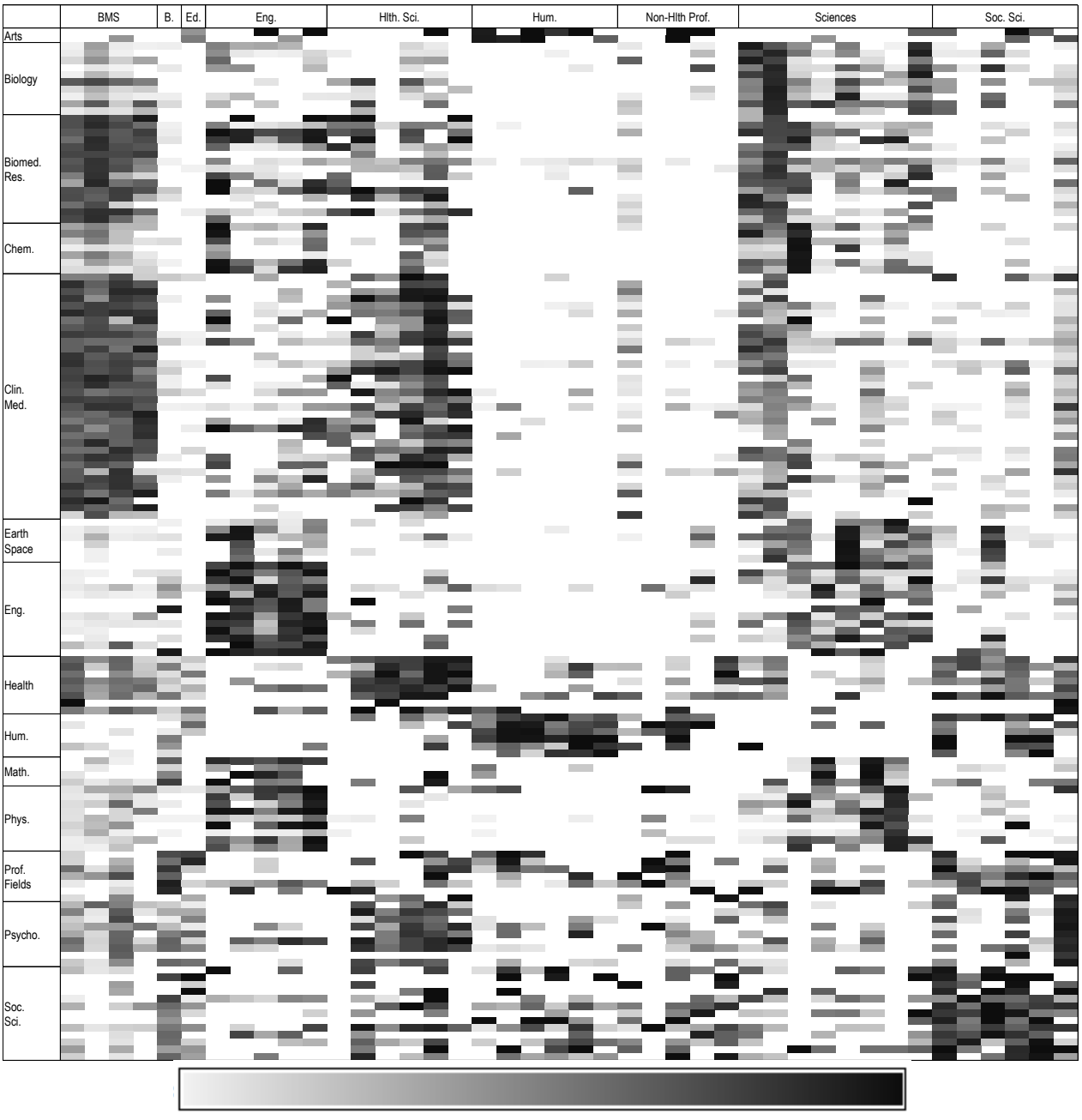
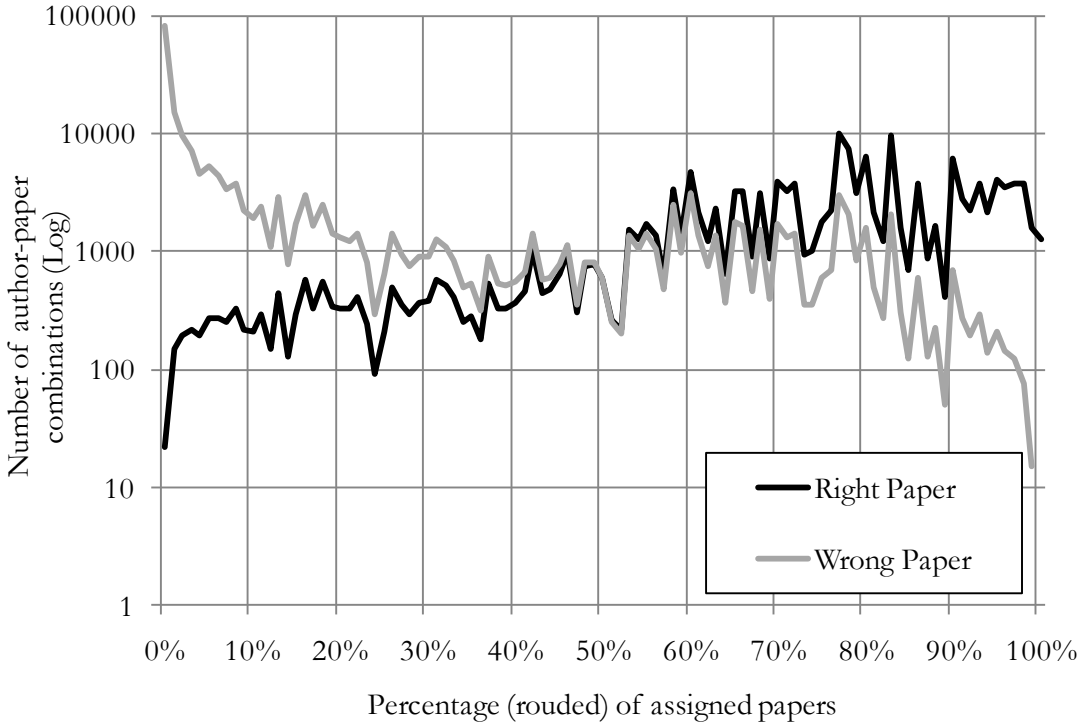


Figure 6. Percentage of papers assigned after manual validation, for each combination of discipline of departments (x-axis) and specialty of publication (y-axis)

Figure 7 aggregates, by rounded percentage of properly attributed papers, the numbers of rejected and of accepted author-paper combinations. We see that the number of wrongly assigned papers drops significantly for department/specialty combinations greater than 80%, and even more after 95%. These percentages were thus used to automatically assign papers in specific discipline that matched researchers

from given departments. In order to reduce the number of false positives, a 95% assignation rate was chosen for papers on which the institution of the author does not appear (but only its province). This includes a total of 17,002 papers, of which 16,518 are properly attributed and only 484 are inaccurately attributed (2.8%). For papers on which the institution of the researcher appears, an 80% attribution rate was used. This attributed 68,785 papers, of which 10.7% were false positives.

Figure 7. Number of rightly and wrongly assigned author-paper combinations, as a function of assignation percentage of papers from a discipline to authors from a department



One must note that all these processes were performed in parallel; a paper assigned with one of these criteria could have been already attributed during another step of the matching. Hence, the numbers of papers presented here include several papers that were already matched using one of the criteria previously presented in this section. Table 4 presents the error rates for all of the steps combined. The inclusion of the algorithm based on the matrix of departments and disciplines of publication added 310 researchers in the subset of those with at least one paper in the core. On the whole, the multiple algorithms used so far for automatically attributed papers for 6,427 researchers, for a total of 50,353 papers and 73,331 author-paper combinations.

Table 4. Overall results of the automatic matching of papers, using core papers, keywords and references previously made, and the matrix of discipline of departments and specialty of papers

Unit of analysis	Manual validation (N)	Automatic assignation		False positives	
		N	%	N	%
Researchers	8 081	6 427	79,5%	610	9,5%
Articles	62 629	50 353	80,4%	1 633	3,2%
Author-paper combinations	97 850	73 771	75,4%	1 750	2,4%

Patterns presented in so far in this paper allow the creation of a dataset of papers that are likely to belong to the right researcher and assign at least one paper to almost 80% of the researchers. They make possible the creation of a core set of papers, as well as of a few other layers of papers, based on the similarity of their characteristics to those included in the core. The following algorithm does the opposite and aims at finding indications that the paper clearly does not belong to the researcher.

As shown on Figure 6, there are several combinations of discipline of departments and specialties of papers where the vast majority of papers were rejected during manual validation. Indeed, if no university researcher from the department X has ever published in the specialty Y, no researcher is likely to do so. Papers falling into these combinations could thus automatically be rejected.

These patterns not only allows the rejection of papers, but also to close researchers' publication files, as all of their papers can either all have been assigned—using the methods previously presented—or rejected using the department/specialty matrix. Using a 50% threshold was optimal, as it automatically rejected 202,928 author-paper combinations, of which 183,656 were real negatives (91%), and only 19,272 were false negatives (9%). These rejected author-paper combinations account for a significant share (90%) of all rejected combinations (226,325).

After all these steps, 5,036 publications files out of the 13,479 (37.4%) were automatically marked as closed (including the 2,256 files for which no paper, either authored by a researcher in the list or by a homograph), as all of their papers were either all assigned or all rejected. Another 6,069 researchers had at least one of their papers automatically assigned (45%), for a total of 50,353 papers, with 1,633 being false positives (3.2%). On the whole, this algorithm provides assignation information on at least one paper for 11,105 (82.4%) out of all 13,479 researchers, or on 8,849 out of the 11,223 researchers (78.8%), when one excludes the 2,256 files for which no paper matched, either authored by the researcher or a homograph. Hence, there are still 2,374 researchers for which no automatic decision on any of their matched papers can be made (assignation or rejection) and, hence, for which a complete manual validation needs to be performed. This algorithm can nonetheless be very helpful, as it automatically assign a large proportion of papers, excludes an even larger one and reduces from 11,223 to 2,374 (79%) the number of researchers for which a complete manual validation has to be performed.

Conclusion

This paper has provided evidence of regularities in researchers' publication patterns, and that these patterns could be used to automatically assign papers to individual and remove papers authored by their homonyms. Two types of patterns were found: 1) at the individual researchers' level and 2) at the level of disciplines.

At the level of individuals, we found that researchers were quite regular in their referencing practices. This could be expected: as shown elsewhere, researchers tend to cite the same material throughout their careers (Barnett and Fink, 2008; Gingras *et al.*, 2008). We thus tested this finding for the subset of Quebec researchers and found that papers with the same surname and initial were always those of the 'right' researcher when at least one of the references of the paper had already been made in one of the papers previously assigned to the researcher. Similarly, researchers also tend to work on the same topics. Using the pool of keywords previously used by researchers and comparing them with papers subsequently published, we found that the use of the same keywords meant in most of the cases that the paper belonged to the same researcher.

At the collective level, two general patterns emerged. The first pattern we found was that the institution of affiliation of a given researcher appeared on most of the papers that rightly belonged to him/her. This simple regularity allowed the creation of a core subset of papers, which could then be used to gather the researchers' other papers using the previous references and previous keywords methods. The other pattern relates to the relationship between the department discipline and the specialty of the journal in which papers are published. For some departments/specialty combinations, a majority of papers belonged to the 'right' researcher, while for other combinations, a majority belonged to homonyms. Thus, the former combinations allowed the automatic assignation of papers, while the latter made automatic rejection of author-paper combinations possible.

Compared with most existing studies on author disambiguation, which were generally performed for a small subset of researchers (Han et al., 2005; Aswani, Bontcheva and Cunningham, 2006; Wooding et al., 2006) or for specific author-article combinations (Boyack and Klavans, 2008) this is an important step forward. That being said, the recent developments in bibliographic databases used in bibliometrics—such as the researcher ID, ORCID, the link between each of the authors and their addresses as well as the indexation of the complete given names of authors—are perhaps even more important, as they are likely to make this assignation easier in the future.

Bibliography

Aksnes, D.W. (2008). When different persons have an identical author name. How frequent are homonyms? *Journal of the American Society for Information Science and Technology*, 59(5), 838–841.

Aswani, N., Bontcheva, K., and Cunningham, H. (2006). Mining information for instance unification, *Lecture Notes in Computer Science*, 4273, 329–342.

Barnett, G.A., and Fink, E.L. (2008). Impact of the internet and scholar age distribution on academic citation age. *Journal of the American Society for Information Science and Technology*, 59(4), 526–534.

Boyack, K.W., and Klavans, R. (2008). Measuring science–technology interaction using rare inventor–author names. *Journal of Informetrics* 2(3), 173–182.

Campbell, D., Picard-Aitken, M., Côté, G., Caruso, J., Valentim, R., Edmonds, S., Williams, G.T., Macaluso, B., Robitaille, J.P., Bastien, N., Laframboise, M.C., Lebeau, L.M., Mirabel, P., Larivière, V., Archambault. É. (2010) Bibliometrics as a performance measurement tool for research evaluation: The case of research funded by the National Cancer Institute of Canada. *American Journal of Evaluation*, 31(1), 66-83.

Cole J.R., and Cole, S. (1973). Social stratification in science. Chicago: University of Chicago Press.

Cota, R.G., Ferreira, A.A., Nascimento, C., Gonçalves, M.A., and Laender, A.H.F. (2010). An unsupervised heuristic-based hierarchical method for name disambiguation in bibliographic citations, *Journal of the American Society for Information Science and Technology*, 61(9), 1853-1870.

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

Enserink, M. (2009). Are you ready to become a number? *Science*, 323, 1662–1664.

Gingras, Y., Larivière, V., Macaluso, B., and Robitaille, J.P. (2008). The effects of aging on researchers' publication and citation patterns, *PLoS ONE*, 3(12), e4048.

Gurney, T., Horlings, E., and van den Besselaar, P. (2012). Author disambiguation using multi-aspect similarity measures. Scientometrics, 91(2): 435–449

Han, H., Zha, H., and Lee Giles, C. (2005) Name disambiguation in author citations using a K–way spectral clustering method. *Proceedings of the 5th ACM/IEEE–CS joint conference on Digital libraries*, pp. 334–343. Available at: http://clgiles.ist.psu.edu/papers/JCDL–2005–K–Way–Spectral–Clustering.pdf

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

Jensen, P., Rouquier, J.B., Kreimer, P., and Croissant, Y. (2008). Scientists who engage in society perform better academically. *Science and Public Policy*, 35(7), 527–541.

Kang, I.S., Seung–Hoon, N., Seungwoo, L., Hanmin, J., Pyung, K., Won–Kyung, S., and Jong–Hyeok, L. (2009). On co–authorship for author disambiguation. *Information Processing and Management*, 45(1), 84–97.

Larivière, V., Macaluso, B., Archambault, E. and Gingras, Y. (2010) Which scientific elites? On the concentration of research funds, publications and citations, *Research Evaluation*, 19 (1), 45-53.

Levin, M., Krawczyk, S., Bethard, S. and Jurafsky, D. (2012). Citation-based bootstrapping for large-scale author disambiguation. *Journal of the American Society for Information Science and Technology*, 63(5), 1030–1047.

Lewison, G. (1996). The frequencies of occurrence of scientific papers with authors of each initial letter and their variation with nationality. *Scientometrics*, 37(3), 401–416.

Merton, R.K. (1973). *The sociology of science: Theoretical and empirical investigations*. Chicago and London: Chicago University Press.

Reijnhoudt, L., Costas, R., Noyons, E., Borner, K., Scharnhorst, A. (2013). Seed+Expand": A validated methodology for creating high quality publication oeuvres of individual researchers. http://arxiv.org/abs/1301.5177

Schreiber, M. (2008). A modification of the h-index: The hm-index accounts for multi-authored manuscripts. *Journal of Informetrics*, 2(3), 211-216

Smalheiser, N.R. and Torvik, V.I. (2009). Author Name Disambiguation, in B. Cronin, Ed. Annual Review of Information Science and Technology, 43,

Torvik, V.I., Weeber, M., Swanson, D.R., and Smalheiser, N.R. (2005) Probabilistic similarity metric for Medline records: A model for author name disambiguation. *Journal of the American Society for Information Science and Technology*, 56(2), 140–158.

Wang, J., Berzins, K., Hicks, D., Melkers, J., Xiao, F., and Pinheiro, D. (2012). A boosted-trees method for name disambiguation. Scientometrics, 93(2), 391-411.

Wooding, S., Wilcox–Jay, K., Lewison, G., and Grant, J. (2006). Co–author inclusion: A novel recursive algorithmic method for dealing with homonyms in bibliometric analysis. *Scientometrics*, 66(1): 11–21.

Zhang, C.T. (2009). The e-index, complementing the h-index for excess citations. *PLoS ONE*, 5(5), e5429

Zuckerman, H.A. (1977). *Scientific elite: Nobel laureates in the United States*. New York: Free Press.

Appendix 1. List of disciplines assigned to journals

Arts
- Fine Arts & Architecture
- Performing Arts

Biology
- Agricult & Food Science
- Botany
- Dairy & Animal Science
- Ecology
- Entomology
- General Biology
- General Zoology
- Marine Biology & Hydrobiology
- Miscellaneous Biology
- Miscellaneous Zoology

Biomedical Research
- Anatomy & Morphology
- Biochemistry & Molecular Biology
- Biomedical Engineering
- Biophysics
- Cellular Biology Cytology & Histology
- Embryology
- General Biomedical Research
- Genetics & Heredity
- Microbiology
- Microscopy
- Miscellaneous Biomedical Research
- Nutrition & Dietetic
- Parasitology
- Physiology
- Virology

Chemistry
- Analytical Chemistry
- Applied Chemistry
- General Chemistry
- Inorganic & Nuclear Chemistry
- Organic Chemistry
- Physical Chemistry
- Polymers

Clinical Medicine
- Addictive Diseases
- Allergy
- Anesthesiology
- Arthritis & Rheumatology
- Cancer
- Cardiovascular System
- Dentistry
- Dermatology & Venerial Disease
- Endocrinology
- Environmental & Occupational Health
- Fertility
- Gastroenterology
- General & Internal Medicine
- Geriatrics
- Hematology
- Immunology
- Miscellaneous Clinical Medicine
- Nephrology
- Neurology & Neurosurgery
- Obstetrics & Gynecology
- Ophthalmology
- Orthopedics
- Otorhinolaryngology
- Pathology
- Pediatrics
- Pharmacology
- Pharmacy
- Psychiatry
- Radiology & Nuclear Medicine
- Respiratory System
- Surgery
- Tropical Medicine
- Urology
- Veterinary Medicine

Earth and Space
- Astronomy & Astrophysics
- Earth & planetary Science
- Environmental Science
- Geology
- Meteorology & Atmospheric Science
- Oceanography & Limnology

Engineering and Technology
- Aerospace Technology
- Chemical Engineering
- Civil Engineering
- Computers
- Electrical Engineering & Electronics
- General Engineering
- Industrial Engineering
- Materials Science
- Mechanical Engineering
- Metals & Metallurgy
- Miscellaneous Engineering & Technology
- Nuclear Technology
- Operations Research

Health
- Geriatrics & Gerontology
- Health Policy & Services
- Nursing
- Public Health
- Rehabilitation
- Social Sciences, Biomedical
- Social Studies of Medicine
- Speech-Language Pathology and Audiology

Humanities
- History
- Language & Linguistics
- Literature
- Miscellaneous Humanities
- Philosophy
- Religion

Mathematics
- Applied Mathematics
- General Mathematics
- Miscellaneous Mathematics
- Probability & Statistics

Physics
- Acoustics
- Applied Physics
- Chemical Physics
- Fluids & Plasmas
- General Physics
- Miscellaneous Physics
- Nuclear & Particle Physics
- Optics
- Solid State Physics

Professional Fields
- Communication
- Education
- Information Science & Library Science
- Law
- Management
- Miscellaneous Professional Field
- Social Work

Psychology
- Behavioral Science & Complementary Psychology
- Clinical Psychology
- Developmental & Child Psychology
- Experimental Psychology
- General Psychology
- Human Factors
- Miscellaneous Psychology
- Psychoanalysis
- Social Psychology

Social Sciences
- Anthropology and Archaeology
- Area Studies
- Criminology
- Demography
- Economics
- General Social Sciences
- Geography
- International Relations
- Miscellaneous Social Sciences
- Planning & Urban Studies
- Political Science and Public Administration
- Science studies
- Sociology

Appendix 2. List of disciplines assigned to departments

Basic Medical Sciences
- General Medicine
- Laboratory Medicine
- Medical Specialties
- Surgical Specialties

Business & Management

Education

Engineering
- Chemical Engineering
- Civil Engineering
- Electrical & Computer Engineering
- Mechanical & Industrial Engineering
- Other Engineering

Health Sciences
- Dentistry
- Kinesiology / Physical Education
- Nursing
- Other Health Sciences
- Public Health & Health Administration
- Rehabilitation Therapy

Humanities
- Fine & Performing Arts
- Foreign Languages Literature, Linguistics & Area Studies
- French/English
- History
- Philosophy
- Religious Studies & Vocations

Non-Health Professional
- Law & Legal Studies
- Library & Information Sciences
- Media & Communication Studies
- Planning & Architecture
- Social Work

Sciences
- Agricultural & Food Sciences
- Biology & Botany
- Chemistry
- Computer & Information Science
- Earth & Ocean Sciences
- Mathematics
- Physics & Astronomy
- Resource Management & Forestry

Social Sciences
- Anthropology, Archaeology & Sociology
- Economics
- Geography
- Other Social Sciences & Humanities
- Political Science
- Psychology