Author Name Disambiguation for Collaboration Network Analysis and Visualization

Authors

Andreas Strotmann
School of Public Health, University of Alberta
Edmonton, Alberta, T6G 2G3, Canada
Email: andreas.strotmann@ualberta.ca

Dangzhi Zhao
School of Library and Information Studies
Edmonton, Alberta, T6G 2J4, Canada
Email: dzhao@ualberta.ca

Tania Bubela
School of Public Health, University of Alberta
Edmonton, Alberta, T6G 2G3, Canada
Email: tbubela@ualberta.ca

In this paper we outline a heuristic algorithm for disambiguating author names of publications via deterministic clustering based on well-defined similarity measures between publications in which their names appear as authors. The algorithm is designed to be used in the construction of a collaboration network, i.e., a graph of author nodes and co-author links. In this context, the goal is to produce a co-authorship graph with network characteristics that are close to those of the "true" collaboration network, so that meaningful network metrics can be determined.

The algorithm we present here is fairly easily comprehended as it does not depend on any sophisticated AI techniques. This is important in the context of policy studies, in which we successfully applied it, as it enables policy makers to judge the soundness of the methodology with considerable confidence. It is also quite fast, making it possible to run large-scale analyses (here, in the order of a hundred thousand publications and in the order of a million names to be disambiguated) on a moderately sized desktop computer within a few days.

The algorithm is, finally, open to improvement via extensions that take into account additional kinds of fields in bibliographic records of publications to provide evidence that two occurrences of similar names belong to the same individual.

Background and Purpose

This paper reports on a method developed for the study of a large-scale Canadian biomedical research network, the Stem Cell Network (Bubela & Strotmann, 2008).

The focus of the study was a practical application of scientometric and network analyses to explore the impact of science funding policies on knowledge flows and collaborative scientific endeavours. In particular, the Stem Cell Network, one of Canada's Networks of Centres of Excellence, is governed both by policies of (1) encouraging a geographically dispersed inter-disciplinary network of most of the country's leading stem cell researchers; and (2) promoting commercialisation of the innovations developed by network researchers. Using a combination of co-authorship network visualization and statistical modeling using various network measures of collaboration intensity as dependent variables, we have explored whether these two policies, facilitating networking on the one hand and commercialization on the other, are synergistic or antagonistic. How embedded are Canadian stem cell researchers within the international research community? Has the Canadian Stem Cell Network facilitated the creation of a geographically dispersed, virtual network of researchers, or are collaboration patterns best explained by institutional affiliation or location? Are Canadian stem cell researchers who commercialise research more or less collaborative when it comes to research publications?

In this context, there are two important considerations. First, we need to perform measurements within a reasonable time frame, and second, the network measures we select as dependent variables in further statistical models need to provide meaningful results, especially since many of the effects we are interested in are quite subtle. This means that we needed to expend considerable effort in adapting scientometrics for application in such real-world policy questions, while balancing effort with meaningful results.

Collaboration Network Construction

We first constructed an author collaboration network from the relevant literature to visualize and then measure the performance of members of our target network. We retrieved the literature from PubMed, a premier source of information for biomedical literature, and processed it into a collaboration network (see below for details).

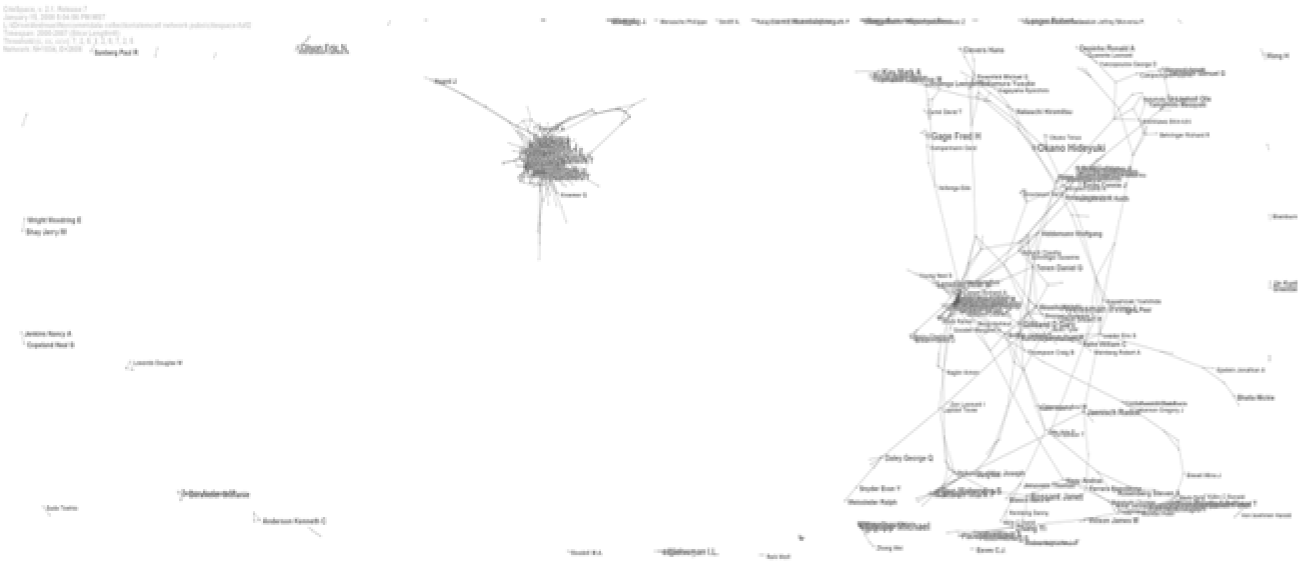

Figure 1 illustrates the reason for the methodology proposed in the present paper. It is a visualization, using Chaomei Chen's CiteSpace (Chen, 2007), of the co-author network contained in the "raw" PubMed file, i.e., using author names exactly as they are reported in PubMed.[1] It exhibits two large clusters and many small ones. Upon closer inspection, we found that the large and compact central cluster in this image and the entire rest of the image both visualize the entire field. The compact cluster comprises authors whose names are listed in PubMed in last-name-plus-initial form, while the rest of the image consists of authors whose names are reported in last-name-plus-full-first-name form.

In many author-based bibliometric studies, author names are normalized to something approximating the latter form – if we based our research policy study on such an author collaboration network, that network would resemble the compact cluster in the center of Figure 1. However, we had strong reason to suspect that the remainder of Figure 1, which identifies authors by their full names, was much closer to a depiction of the "real" co-author network than the compact cluster. In particular, we found that Asian names like “Wang, X” dominated in the compact cluster, while in the rest of the graph, many dozens of full author names corresponded to such last-names-plus-initials, and their dominance was greatly reduced.

Given our goal of extracting high-quality collaboration intensity measures from a collaboration network in order to test a subtle effect for policy analysis, Figure 1 taught us that we could not avoid tackling one of the so far largely unresolved issue of bibliometrics, namely, *author name disambiguation*, i.e., the identification of individuals from author names, or equivalently, the identification of author oeuvres within a literature.

Collaboration network studies

By contrast, collaboration network studies in the scientometrics literature have so far largely been restricted in practice to relatively small-scale co-author network studies in which full author names are reported (Liu &al., 2005; Yin & al., 2006). Liu & al. (2005) provide an excellent literature review of co-author network studies, but this is not the central theme of the current paper. It is not clear in these studies, however, if or how the author names were cleaned for analysis, but based on the data set sizes reported (the largest listing 5,000 authors), manual cleaning is quite feasible if the underlying dataset provides full names of authors already.

Other studies of coauthor networks have relied on high-quality hand-compiled networks such as that underlying the famous Erdős number project. In these cases, no disambiguation was necessary. Finally, a series of seminal papers by Newman (2001a; 2001b) studied large-scale coauthor networks without cleaning the data beforehand, a problematic approach as he was the first to admit, but admissible in the context of estimating global network statistics for large-scale networks.

As Figure 1 illustrates, we were not able to avoid author name disambiguation, and with a dataset approximately a factor thousand larger than those used in most previous coauthor network studies, hand cleaning was impractical.

Author Name Disambiguation

Author name disambiguation is one of the great unresolved issues in bibliometrics (Andrade & al., 2006). The two major citation indexes both have recently added author identification to their databases, although their proprietary identification algorithms tend to be quite cautious about assigning identity (Thomson Reuter Science, 2009; Scopus, 2009).

Kang & al. (2008) provide an excellent review of the problem and of the literature that explores approaches for dealing with it. We will briefly summarize it here, and then focus on the part of the bibliometrics literature that is most relevant to the present study.

As Kang & al. (2008) point out, the author name disambiguation problem consists of two independent sub-problems, which we may term *collation* and *identification*. The collation subtask is to identify all possible different variations of a name that an individual author may be listed under in a data source. There is quite some literature available on this problem, but we will not go into it as we do not address this aspect of author name disambiguation here, except in a fairly rudimentary fashion.

The identification subtask, which this paper focuses on, takes as input a set of author name occurrences that all potentially refer to the same individual, and attempts to solve the problem that more than one individual may be represented in this set of name occurrences. The goal is therefore to identify all the individuals named in this set, and to assign each occurrence of a name in this set to one of the individuals identified. This is usually equivalent to identifying an individual author and that author's oeuvre in a citation database.

Andrade & al. (2006) urge the creation of what amounts to authority files for author identification in the science literature, which unequivocally identify a name with an individual and would usually be maintained manually (and, they suggest, could be maintained in a mass collaboration).

However, these authority files do not exist, yet. Author-based scientometrics studies are therefore always hampered by the fact that the same name in different places may not always refer to the same individual. Classic author-based methods such as author co-citation analysis therefore do not so much analyze authors as they analyze author pairs, and White & McCain (1998) argue convincingly that this dramatically reduces the author name ambiguity problem.

For other author-based bibliometric methods this is not true, as we show below, and the author disambiguation problem needs to be tackled in order to proceed with a meaningful scientometric study. A few studies have therefore looked into the identification subtask of author name disambiguation.

Torvik & al. (2005) created a web service that will report, for any given record of the PubMed database and one of its authors, a ranked list of PubMed records most likely to belong to the same oeuvre. This service can be used to identify the oeuvre of individual authors, but it leaves it to its users to apply a good cut-off value for the similarity metric at which to separate oeuvre from non-oeuvre. The algorithm is based on an unsupervised machine learning algorithm, fed with information extracted from each record in PubMed using, among others, text processing on several fields of the record. The method is applied to the entire PubMed/Medline database of roughly 15 million records, and is computationally expensive.[2]

Very recently, Kang & al. (2008) reported on their experiments with a series of algorithms for determining author identity, relying exclusively on collaboration patterns between authors to drive the disambiguation. Unlike Torvik's method, they do not rely on a machine learning algorithm. They report success rates up to 80% for the author identification task using only bibliographic information, and up to 85% if using manually added information on "implicit" collaborations between authors, in a Korean-only dataset of about 9000 documents in a highly collaborative research field.

In the present paper, we report on an extension and variation of the algorithms discussed by Kang & al., applied to a large dataset extracted from PubMed, which relies exclusively on bibliographic information for the author name disambiguation task. It was developed and run on a standard modern desktop computer, where a full study of more than 100,000 document records containing close to a million author name occurrences took just a few days.

Author identification is a problem that is also being discussed in the computer science and digital libraries literature (e.g. Han, Zha, & Giles, 2005). As in Torvik's (2005) work, the basic idea in that literature appears to be to use supervised or unsupervised machine learning algorithms to address this issue, frequently using co-author information as input. The results generally reported in that literature are not impressive (in the order of 50% success, although the success measure reported there cannot be compared to the one we report below). Papers like Han's (2005), moreover, utilize algorithms that require pre-knowledge of the number of distinct individuals in each group of comparable author names – clearly an untenable proposition in our context. These more theoretically interesting studies had therefore clearly not matured yet to the level that it would have been possible for science policy researchers like us to apply their results with sufficient confidence in practice at the time we ran our analyses.

Methods

Data Collection

The collaboration network we aimed to construct was based on a dataset of roughly 160,000 PubMed records, with roughly six coauthors per paper on average. The data collection methodology that we developed to construct this dataset is described in some detail in a companion paper (Strotmann & al., 2009). We briefly summarize it here.

The administrator of the Canadian Stem Cell Network provided us with a then recently compiled list of about 1,400 works funded by the network and a list of all current and former primary investigators (PIs) of the network (more than 90).

The list was converted to a search strategy for the Scopus database using a simple script we wrote. In blocks of about 100 each, the search strategy was run in Scopus, and the result downloaded as full records in CSV format. For each such block, the citing literature for that set of results was also determined in Scopus and downloaded in the same format.

This resulted in about 500 unique records of publications funded by the network, and about 7000 total unique full records downloaded from Scopus as citing papers.

We used more complex software to parse this dataset into an appropriate data structure which contained a broken-down record of the relevant components of each cited reference indexed by Scopus for these records (on average, each of the 7000 citing document contained about 60 cited references).

From this data structure, we created a series of search strategies for the PubMed Batch Citation Matcher which succeeded in matching 100% of citing documents and 90% of the cited references in the data set. The results returned by the citation matcher were then converted into a search strategy for PubMed itself in order to retrieve the matched documents (both citing and cited) from PubMed in XML format. This dataset, comprising about 160,000 full PubMed XML records, formed the dataset for this study. Since there were on average six coauthors per paper in this dataset, the dataset contained close to a million author references that needed to be disambiguated for our study.

Data Preparation

In preparation for the subsequent author name disambiguation, the PubMed XML records were parsed and separated out into a number of tables, each indexed by a record's PubMed ID. We created separate tables for (a) titles; (b) coauthors; (c) journal or source name; (d) year of publication; (e) volume of publication; (f) number of publication; (g) page number(s) of publication; (h) list of major MeSH codes assigned by PubMed to this publication. These tables could be loaded individually or collectively as required into a program running the author name disambiguation algorithm detailed below.

Author Name Disambiguation Algorithm, Phase 1

In the first phase of the disambiguation process, our algorithm classifies a set of name occurrences as referring to a single individual if:

1. All author names for this individual are mutually compatible.
2. There is positive evidence that indicates that all occurrences in the set refer to the same individual.
3. There is more positive evidence for this particular choice than for potential

alternatives.

The rest of this section explains in more detail the algorithm that implements this idea.
In this algorithm, author names are considered compatible if:

1. Their last names and first initials reduce to identical ASCII sub-sequences after Unicode compatibility decomposition normalization (referred to below as normalized and reduced).
2. Both sequences of first names consist of compatible first names in the same order, where two first names are compatible if:
    1. One of them is empty (i.e., the corresponding sequence of the other name is a compatible extension of this one); or if
    2. One of them is an initial and that initial normalizes and reduces to the first letter of the reduced and normalized other first name (which may be an initial, too); or if
    3. Neither first name is an initial or empty (i.e., they are both full first names) and they both normalize and reduce to the same sequence of ASCII characters.

A number of different types of evidence are considered as positive evidence that two occurrences of similar names refer to the same individual:

1. Equivalence (as opposed to compatibility) of full names, where equivalence is determined by (a) separately normalizing and reducing last names and first name sequences, (b) removing special characters from these reduced name parts, and (c) requiring equality of the resulting sequences.
2. The names occur in two (different) papers whose respective coauthor lists have more than one member in common (i.e., at least one member each beyond the two names under consideration). "In common" means that the names of those coauthors are compatible in the above sense.
3. There are common topics covered by the two papers in which the two similar names occur:
    1. The papers appear in the same journal; or
    2. The papers share a common major MeSH code assigned by PubMed.

Finally, when considering which of several potential candidate individuals to assign a particular name occurrence to, or when considering whether to consider two previously separate individuals as one, a degree of similarity between two individuals, i.e., between two groups of occurrences of similar names, is computed as follows:

1. For each of the different types of positive evidence described above, a list of common features between the two groups is compiled (e.g., common co-authors or common journal names). For a group of occurrences, the union of features of all members of that group is used.
2. Each of the common features is then weighted by the maximum number of times that that feature occurs in one of the groups, since a feature that occurs many times in a group may be considered characteristic of that group, and for two groups to have such a characteristic feature in common is strong evidence that they should be considered one group.
3. Finally, the similarity measure is computed by summing up the weights of all common features.

With these concepts in place, the first phase of the algorithm iterates over all the different normalized and reduced last-name-plus-first-initial combinations that occur in the data set. For each such combination, the algorithm assigns each and every occurrence of an author name that normalizes and reduces to this combination to one of possibly several individuals, as follows:

1. Initially, those groups of name occurrences that normalize and reduce to identical full names (i.e., to names that contain full first names rather than only initials as first names) are considered separate individuals, while initials-only name occurrences are left in a separate list for later processing.
2. For each of these initial individuals whose names contain initials, we pair this individual with another individual (if present) whose name is equivalent except for the initial, as these two individuals are likely candidates for merging into one. If no such pairings exist, and the list of initials-only name occurrences is empty, the result of the first step is considered the final result of the algorithm.
3. For each (“stripped”) individual that is paired in step 2 with (“expanded”) individuals that have additional name components, we attempt to distribute that individual’s oeuvre over the expanded individuals. For each name occurrence of the “stripped” individual name in its oeuvre, therefore, we

   1. Compute that name occurrence’s similarity to each of the “expanded” individuals.
   2. If there is an “expanded” individual (i) with positive evidence (in the sense above) that the name occurrence might be that of the expanded individual’s; (ii) containing only compatible name occurrences; and (iii) with stronger positive evidence than any other expanded individual (if any), the “stripped” name occurrence is moved to that expanded individual.

4. We now process the list of initials-only name occurrences of step 1, using the same method as in step 3, but with the initials-only name occurrence as a “stripped” individual which is paired with each one of the individuals currently under consideration in this iteration.
5. For those initials-only name occurrences that could not be mapped to full name individuals in the previous step, we build a graph of author name occurrences connected by an edge whenever there is (any) positive evidence that these two occurrences refer to the same individual. The connected components of this graph are then each considered as separate individuals to be added to the set of individuals as above.

Author Name Disambiguation Algorithm, Phase 2

Preliminary explorations of the resulting collaboration network showed that the clustering algorithm that we applied above was overly pessimistic, separating known authors into several “individuals”. We therefore decided to balance the excess of false negatives (i.e., missed identifications of individual clusters) at the expense of possible false positives (i.e., spurious identifications of two distinct individuals as one) in the hope of gaining a network with statistical properties that are more like those of the (unattainable) “correct” network of individuals and their collaborations.

In this phase of the algorithm, we rely exclusively on collaboration links: if two individuals have similar names and have at least one collaborator in common, but have never collaborated with each other, they are likely a single individual.

For all sets of author names with identical normalized and reduced last name and first initial, we therefore

1. Construct a graph that links two individuals whenever they have a common co-author;
2. Determine the non-trivial connected components of this graph; and
3. For each connected component, merge its member individuals into a single individual – unless that individual would end up being his or her own co-author by doing so.

This step is repeated twice instead of repeating it until it converges, in an attempt to keep a balance on potential false-positive errors potentially introduced through this method.

Findings

Collaboration Network

The final model of the knowledge network embedded in the data we collected (records of roughly 160,000 publications) identified 361,064 distinct individual authors. On average, a paper had six coauthors; and an individual co-authored 2.7 papers.

There were 2.77 million pairs of collaborators in this network. On average, an author collaborated with eight distinct individuals. We conclude that on average, only about a third of an author's collaborators on a given paper are distinct from that author's collaborators on other papers, while the other two thirds tend to be in common with other papers by the same author.

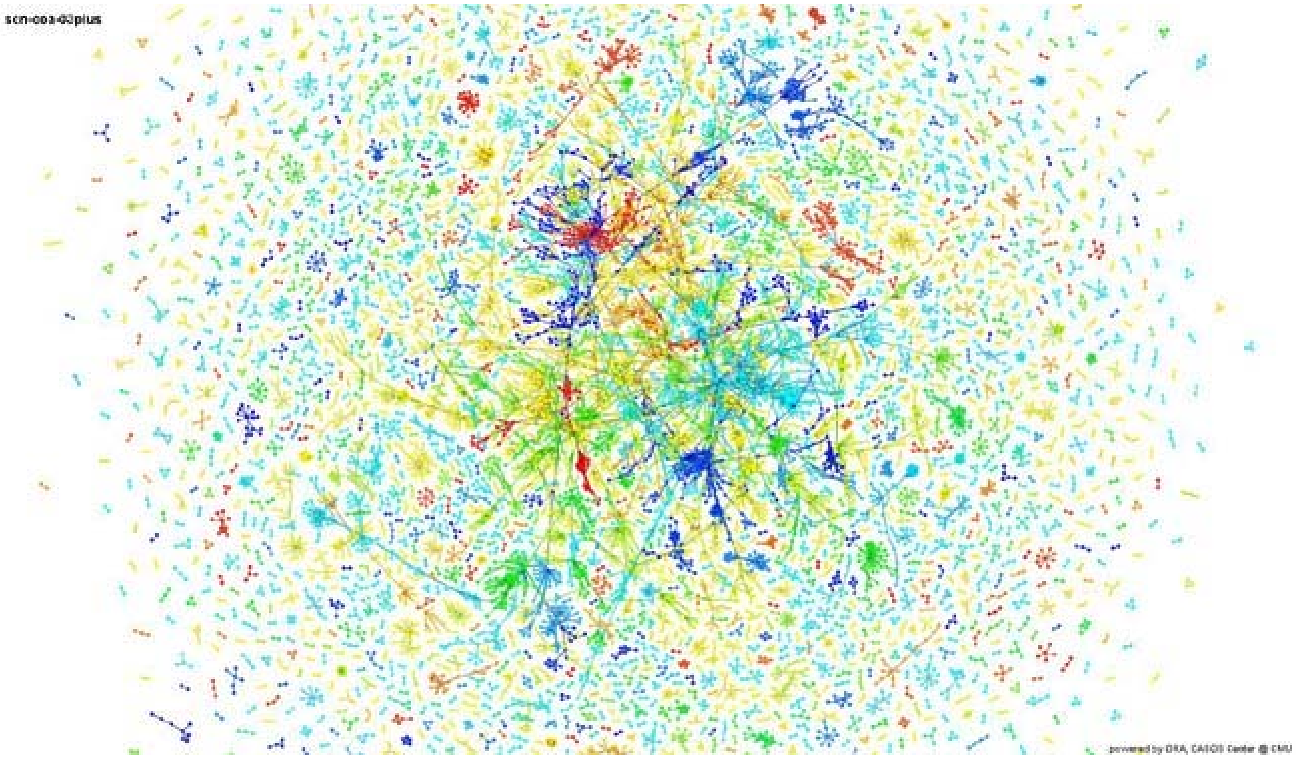

Figure 2 shows a visualization of a much reduced version of the collaboration network constructed from the author disambiguation reported here. Links between collaborators are shown only if the two individuals have collaborated on at least twelve publications, and individuals are shown only if they have collaborated with at least one individual at least twelve times.[3]

The central components of this visualization show a similarly filigreed structure as the full-name part of Figure 1 rather than the compact, seemingly almost fully connected structure of the initials-only core of Figure 1. It therefore appears that our algorithm provided us with a reasonable approximation of the “true” collaboration network that produced the literature we analyzed here.

The following section studies the success of our method using a more detailed approach.

Author Name Disambiguation

The research network we analyzed in this case study has had 96 individual PIs over its history, which at the time of data gathering spanned about five years. In order to judge how well the algorithm fared, we identified all the individuals in the constructed knowledge network that corresponded to these network PIs.

Of the 96 network PIs, two were unidentifiable based on the information we had on them. Of the remaining 94 PIs,

- 76 were correctly identified as individuals by this algorithm;
- 1 had changed her name in marriage and was identified as two separate individuals, each correctly identified;
- 2 had frequently published using two different first initials (they both used their second first name as their main first name). Both factored out into two separate individuals, each correctly identified.
- 15 PIs always published under the same name, but nevertheless separated out into more than one individual in this algorithm, among them
    - social scientists who were PIs in this largely bio-medical science network; and
    - 2 biomedical scientists whose commercialization agenda made them outliers in other parts of our study, too.

Since we purposely avoided taking into consideration name variations that did not share the same normalized and reduced last name and first initial, we calculate our success rate as 76 out of 91 or as 82 out of 94 (since these three individuals' two identities each were each correctly identified).

Discussion

Clearly, therefore, there is room for improvements to our approach to author identification, although the success rate is similar to the best results others have reported in the literature (Kang & al., 2008).[4]

It is therefore instructive to take a closer look at the 15 PIs whose oeuvres were identified as belonging to more than one individual.
The first thing we see is that the algorithm failed miserably for the social scientists in this interdisciplinary research network (mostly law, ethics, and policy researchers). Most of them factored out into more than one individual using our approach. Indeed, this group accounts for 40% of the error cases.

There are a number of possible explanations for this. In the data collection phase, for example, we may have struck a database bias in PubMed against the social sciences (after all, it is primarily a database of medical research). There might also be a lack of citation information in Scopus for law research publications, which follow radically different citation styles from other sciences, after all, and may thus have been left out of the automatic citation indexing process employed by Scopus – most of the social scientists in this group are law researchers or publish in part in law journals.

Field differences in collaboration patterns, which this algorithm crucially depends upon, are another likely explanation. In the social sciences the average number of coauthors of a paper has been reported to be less than two in some areas (Zhao & Strotmann, 2008), less than a third of what we found in this bio-medically dominated field. The algorithm may thus simply not have sufficient information available to identify an individual social scientist based on collaboration patterns, which may explain why all successful author disambiguation studies we found in the literature reported applications in highly collaborative research areas.

The latter explanation may also serve for the two bio-medical researchers with by far the strongest commercialization agenda in the group, as measured by the ratio of the number of patents they applied for world-wide and the number of science journal publications they coauthored. With these researchers, the need to cleanly separate their roles as public scientists on the one hand and as private entrepreneurs on the other, each presumably with their own separate labs and funding structures, may have conditioned them towards compartmentalization of their research activities, with the result that their collaborators would partition into non-overlapping sub-groups, which would defeat the reasoning behind the collaboration-pattern based author name disambiguation algorithm. Indeed, the only network PI who applied for more patents than he published papers, a highly prolific researcher, separated out into half a dozen “individuals” using our method, much more than anyone else.

Finally, we noticed that several of the remaining network scientists who fragmented into two or more “individuals” in this algorithm had had a long and varied career behind them when they became members of the research network we studied. In addition to their current research, they tended to be cited for some of their older, seminal research papers, which resulted in sometimes sizable gaps in the timeline of coverage of their oeuvres in our dataset. In these cases, the algorithm was unable to bridge these gaps whenever they grew too large, and the authors fragmented into several “individuals” based on time slices; with a more complete dataset of publications, it is likely that most of these “individuals” would have merged into one.

Nevertheless, the success rate of the algorithm is 83-87% for the full list of PIs, and reaches more than 90% for the biomedical portion of the network members. The resulting collaboration network visualized in Figure 2 shows the kind of network structure that our initial forays led us to expect (Figure 1). Thus, with the exception, unfortunately, of the social science areas of the network, we expect that the algorithm presented here allowed us to create quite a close approximation of the collaboration network represented by the literature collected. For the purpose of our subsequent statistical analysis of characteristics of PIs using collaboration network measures as one set of parameters, we therefore removed the social scientists from consideration. The final analysis of collaboration and other behaviours of the remaining network PIs produced excellent results, to be reported elsewhere (see (Bubela & Strotmann, 2008) for a preliminary report).

The success rates we can report for the algorithm are quite close to those reported by others in the literature (Kang, 2008). For the core biomedical research within the network, the success rate we report (greater than 90%) exceeds those previously reported in the literature, despite solving a slightly more complex problem (Kang et al tested a dataset with only full author names) that relies exclusively on bibliographically available information (Kang & al. achieve success rates comparable to ours only when “implicit” collaboration information determined from manual web searches is added to their data). This is likely due to the fact that we do not rely exclusively on collaboration information, but also on other types of evidence that indicates that a publication belongs to an individual author’s oeuvre.

Torvik & al. (2005) report higher success rates on similar data, namely, PubMed records. While we performed our analysis on about 160,000 documents and their authors, their study used the entire PubMed database of about 15 million entries as a dataset. As noted above, we expect that our algorithm would have performed significantly better as well had it been applied to a more comprehensive dataset. Their study apparently required supercomputing resources to perform, both due to the size of the dataset it was performed on, and (presumably) the nature of the algorithm they employed. Quite sophisticated text analysis appears to have played a major role in their study, unlike ours, which explicitly did without any type of text analysis due to lack of computing power. Finally, their algorithm crucially depends on an unsupervised machine learning algorithm, which may make it harder for policy makers to judge the adequacy and suitability of their results and underlying assumptions.

Conclusions

We reached our primary goal for this algorithm, namely, to be able to perform some quite subtle network measurements with a fairly low error rate, at least by scientometric standards. To our knowledge, this is the first time that author disambiguation was implemented specifically to create the co-author network model for a large-scale science policy research study.

We also reached our secondary goals. The algorithm is reasonably straightforward, and allows both those who perform policy studies and those who implement policy to judge its soundness and adequacy for their purposes with some confidence. It runs in a reasonable time on a reasonably equipped desktop computer – any standard laptop or desktop currently being offered at retail stores can perform the entire analysis we report here within a few days. With minor modifications, especially with respect to the use of MeSH terms as evidence for author identity, the algorithm should be applicable to data sources other than PubMed.

However, the algorithm as it stands fails when applied to the social science part of the literature. While this may in part be due to a problem in our data gathering phase (PubMed collects medical, not social science literature, which meant that that part of the research field might have been severely underrepresented in our study), we suspect that it is also due to a difference in field characteristics, with social scientists both collaborating and publishing significantly less than biomedical scientists, resulting in insufficient evidence available for reliable author identification.

It would therefore be worthwhile to explore a wider range of bibliometric evidence when identifying individual authors. The algorithm as it stands is quite open to such extensions, be they self-citations or text similarities.

Nevertheless, there remain unresolved issues in author oeuvre identification that are not covered at all here nor in any other study we are aware of, especially with respect to the wide variety of phenomena that lead to an individual author's publishing under different names.

We have found that author collaboration network analysis can provide significant insights into research policy questions (Bubela & Strotmann, 2008), and that the correct identification of individual authors in the network is the single greatest stumbling block for such a study (affiliation identification being the second largest). As our final conclusion, we would therefore like to second others (Andrade & al., 2006) who have recommended the creation of reliable authority files for the identification of individual authors (and, we may add, institutions). We can add that the type of algorithm that we employed here may help making this a feasible task, in that it can provide a first draft of reasonable quality of such an authority file.

Acknowledgements

This research was funded in part by the Social Sciences and Humanities Research Council of Canada (grants 512 2003 1004 and 410 2008 02530), by Genome Prairie (grant PP780315) and by the Stem Cell Network, a Canadian Network of Centres of Excellence (grant PP52). Figures 1 and 2 have been taken, with permission through a Creative Commons License, from Bubela & Strotmann (2008).

References

Andrade, M. & al. (2006), *Workshop on scholarly databases and data integration*. Bloomington, Indiana. Retrieved August 8, 2008, from http://scimaps.org/meeting_060830.php.

Bubela, T. & Strotmann, A. (2008). Designing metrics to assess the impacts and social benefits of publicly funded research in health and agricultural biotechnology. Case study, *The International Expert Group on Biotechnology, Innovation and Intellectual Property*. Retrieved January 26, 2009, from http://www.theinnovationpartnership.org/data/ieg/documents/cases/TIP_Innovation_Metrics_Case_Study.pdf

Chen, C.M. (2007). CiteSpace: visualizing patterns and trends in scientific literature. Retrieved November 6, 2007, from http://cluster.cis.drexel.edu/~cchen/citespace/

Han, J., Zha, H.Y., & Giles, C.L. (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*.

Kang, I.S. & al. (2008). On co-authorship for author disambiguation. *Information Processing and Management.* In press, doi:10.1016/j.ipm.2008.06.006

Liu, X.M. &al. (2005). Co-authorship networks in the digital library research community. *Information Processing and Management*, 41 (2005), 1462-1480

Newman, M.E.J. (2001a). Scientific collaboration networks: I. Network construction and fundamental results. *Physical Review E*, 64, 016131

Newman, M.E.J. (2001b). Scientific collaboration networks: II. Shortest paths, weighted networks, and centrality. *Physical Review E*, 64, 016132

Scopus (2009). Scopus Author Identifier. Retrieved June 5, 2009, from http://info.scopus.com/authoridentifier/.

Strotmann, A., Zhao, D., & Bubela, T. (2009). A multi-database approach to field delineation. To appear in *12th International Conference of the International Society for Scientometrics and Informetrics*, 2009, Rio de Janeiro, Brazil

Thomson Reuters Science (2009). Distinct Author Identification System. Retrieved June 5, 2009, from http://science.thomsonreuters.com/support/faq/wok3new/dais/.

Torvik, I.V, & al. (2005). A probabilistic similarity metric for Medline records : a model for author name disambiguation. *Journal of the American Society for Information Science and Technology*, 56(2), 140-158

White, H. & McCain, K.W. (1998). Visualizing a discipline: an author co-citation analysis of Information Science 1972-1995. *Journal of the American Society for Information Science*, 49 (4), 327-355

Yin, L.C. & al. (2006). Connection and stratification in research collaboration: an analysis of the COLLNET network. *Information Processing and Management*, 42 (2006), 1599-1613.

Zhao, D.Z. & Strotmann, A. (2008). Comparing all-author and first-author co-citation analyses of Information Science. *Journal of Informetrics*, 2(3), 229-239

Footnotes

Footnote 1. CiteSpace attempts to reduce a huge and complex network such as the one described here to core features and core links, and visualizes the reduced network in such a way that these features become visible. Here, we used its author collaboration network visualization feature to illustrate the results of a typical network analysis of a “raw” collaboration network.

Footnote 2. At least too expensive, apparently, to re-run the analysis to incorporate updates to the underlying dataset at regular intervals, according to their website.

Footnote 3. The visualization was created by (a) exporting the full collaboration network created in this study to a file in Pajek's .net format; (b) importing the file into *ORA; (c) performing the above-mentioned network reductions in *ORA; and (d) visualizing the resulting network in *ORA using node coloring by its implementation of Newman grouping. (Pajek and *ORA are both network analysis and visualization tools.)

Footnote 4. This comparison of success rates has to be taken with a large grain of salt, of course. The Kang & al. success rates are highly accurate, as they rely on comparing algorithm outcomes with manual disambiguation results for thousands of authors, whereas we can only report preliminary estimates of the accuracy of our algorithm based on manual checking of results for a fairly small and non-random sample of less than 100 out of the hundreds of thousands of authors we disambiguated.