JISTaP http://www.jistap.org
Journal of Information Science Theory and Practice

Research Paper
J. of infosci. theory and practice 2(2): 06-15, 2014
http://dx.doi.org/10.1633/JISTaP.2014.2.2.1

The Impact of Name Ambiguity on Properties of Coauthorship Networks

Jinseok Kim
Graduate School of Library and Information Science, University of Illinois at Urbana-Champaign, USA
E-mail: jinseok.kim.uiuc@gmail.com

Heejun Kim
School of Information and Library Science, University of North Carolina at Chapel Hill, USA
E-mail: heejunk@email.unc.edu

Jana Diesner *
Graduate School of Library and Information Science, University of Illinois at Urbana-Champaign, USA
E-mail: jdiesner@illinois.edu

ABSTRACT
Initial based disambiguation of author names is a common data pre-processing step in bibliometrics. It is widely accepted that this procedure can introduce errors into network data and any subsequent analytical results. What is not sufficiently understood is the precise impact of this step on the data and findings. We present an empirical answer to this question by comparing the impact of two commonly used initial based disambiguation methods against a reasonable proxy for ground truth data. We use DBLP, a database covering major journals and conferences in computer science and information science, as a source. We find that initial based disambiguation induces strong distortions in network metrics on the graph and node level: Authors become embedded in ties for which there is no empirical support, thus increasing their sphere of influence and diversity of involvement. Consequently, networks generated with initial-based disambiguation are more coherent and interconnected than the actual underlying networks, and individual authors appear to be more productive and more strongly embedded than they actually are.

Keywords: bibliometrics, name ambiguity, initial based disambiguation, coauthorship networks, collaboration networks

1. INTRODUCTION

Authorship data are not only used to evaluate individuals for employment, tenure, and funding, but also to understand fundamental principles of scientific collaboration, communication, and productivity. Thus, scholars as well as organizations involved with the progress, promotion, and management of science

Open Access

Accepted date: June 21, 2014
Received date: June 9, 2014

***Corresponding Author:** Jana Diesner
Assistant Professor
Graduate School of Library and Information Science
University of Illinois at Urbana-Champaign, USA
E-mail: jdiesner@illinois.edu

have a strong interest in gaining a better, actionable understanding of these processes (Torvik, Weeber, Swanson, & Smalheiser, 2005). A known but insufficiently solved problem in this domain is name disambiguation, i.e. identifying whether a set of name strings refers to one or more real-world persons. This task can be very difficult, especially when an author's identity is only represented by a string of characters. For example, when encountering spellings of seemingly similar names, such as 'Smith, Linda' and 'Smith, L.', it is not always clear whether these names represent the same person or not. The given problem can get more complicated, especially when people use different names, e.g. due to marriage, translating their name into another language, or inconsistent use or spelling of names.

One solution to solve name ambiguity is to manually inspect bibliometric data. For example, two name instances that appear in two different citation records can be evaluated for identity by considering additional information, e.g. people's web pages or curricula vitae, as well as meta-data on the publication, including keywords and index terms. The caveat with this approach is its limited scalability and related costs. Consequently, manual verification can hardly be applied to large datasets that contain thousands or millions of name instances.

A more scalable solution is computational approaches that consider attribute data. To achieve high accuracy with this approach, scholars have typically used a two-step process (Treeratpituk & Giles, 2009): First, a pair of names appearing in bibliometrics records are compared to each other based on attributes of the author(s), e.g. the surface form of their name as well as their affiliation, and of the paper, e.g. its title and the title of the journal. These pairwise comparisons produce similarity profiles between any pair of name instances, where distance is determined based on rules as well as metrics such as edit-distance functions. The similarity profiles are then used to make a binary decision per pair ('yes' for matched names or 'no' for unmatched names) or to calculate some (probabilistic) similarity score between 0 and 1. In a second step, the authors' names are clustered based on the decision or score of pairwise comparisons. For more details on this procedure, we refer readers to Smalheiser and Torvik (2009).

Yet another automated approach in this field is a solution based on heuristics that are employed to assign identities to name instances based on one or more parts of a name string. Regular expressions, which identify morphological similarities and different types of congruence on the surface-form level between any pair of name instances, are commonly used for this purpose. For example, if two name instances share the same last name and same first name initials, these two names can be assumed to refer to the same author (e.g., Newman, 2001). In fact, this kind of approach has become a dominant name disambiguation strategy in coauthorship network research (Strotmann & Zhao, 2012). A major reason for the wide adoption of this strategy is that name ambiguity has been supposed to have small to moderate impact on the resulting network data and analysis results (Barabasi et al., 2002; Milojević, 2013; Newman, 2001). This assumption has been insufficiently investigated. We herein fill this gap and complement prior work on this issue by comparing the statistical properties of coauthorship networks constructed from a) a proxy for ground truth data, more specifically from the DBLP Computer Science Bibliography (DBLP hereafter) (Ley, 2002) and b) network data built from the same dataset, but after applying initial based disambiguation to it.

In the following, we first review how errors or biases induced by initial based name disambiguation have been addressed in prior work. Then, we empirically estimate the impact of initial based disambiguation on network data and findings. We conclude with a discussion of research implications.

2. BACKGROUND

The following three approaches to name disambiguation based on initials and last names have been suggested for coauthorship network research (Milojević, 2013). First, one could rely on any given author's first name initial plus their last name. This is also known as the "first initial method." With this approach, matches in the last name and the initial of the first name are regarded as referring to the same person, regardless of the existence of or differences in middle name initials. The second approach considers the initials of the first and the middle name. This is also known as the "all initial method." Here, matches in first and middle name initials and in last names are assumed to represent the same person. The third way is a hybrid method, which

uses the first initial method as a baseline. Then, if a name entailing a first name initial and a last name has two or more potential match candidates with names entailing different middle name initials, all potential match candidates are considered as different identities.

Table 1. Illustration of Types of Initial Based Disambiguation

Methods	Examples from Milojević (2013)	Decision
First Initial Method	Jackson, P. Jackson, P. A. Jackson, P. S.	all the same author
All Initial Method	Jackson, P. Jackson, P. A. Jackson, P. S.	three authors
Hybrid Method	If 'Jackson, P.' has TWO or more match candidates with different middle name initials, Jackson, P. Jackson, P. A. Jackson, P. S.	three authors
	If 'Jackson, P.' has only ONE match candidate with a middle name initial, Jackson, P. Jackson, P. A.	all the same author

To gain a better understanding of which of these methods has been used or studied in bibliometrics research, we screened 298 articles that contain the term 'co-author' or 'coauthor' in the title, abstract, or keyword section in eight journals from 1978 to 2013. We considered the following journals: *Information Processing & Management, Journal of Information Science, Journal of Informetrics, Journal of the American Society for Information Science and Technology, Physica A, Physical Review E, Plos One*, and *Scientometrics*. About 70% of the retrieved articles were focused on studying a) the number of authors per paper to analyze trends over time or b) coauthoring across institutions and nations (e.g. Leydesdorff & Sun, 2009). Both applications do not require author name disambiguation in most cases. However, name disambiguation is needed for the remaining 30% of papers, where coauthorship networks are analyzed in which node names consist of last names and first- and/or middle-name initials (e.g., datasets from Web of Science or Scopus). In some of these papers it is clearly indicated that they used the first initial method (9 papers, e.g. Bettencourt, Lobo, & Strumsky, 2007; Liben-Nowell & Kleinberg, 2007) or all initial method (4 papers, e.g. Milojević, 2010; Newman, 2001). Only one paper disambiguated with the hybrid method (Yoshikane, Nozawa, Shibui, & Suzuki, 2009). Some scholars just indicated that an initial based disambiguation had been performed without details on the strategy (6 papers, e.g. Barabasi et al., 2002; Fiala, 2012; Lee, Goh, Kahng, & Kim, 2010; Rorissa & Yuan, 2012). Several others clearly stated that they did not resolve name ambiguities at all but relied on full surnames and initialized given names to identify authors (e.g. Braun, Glanzel, & Schubert, 2001; Lariviere, Sugimoto, & Cronin, 2012; Wagner & Leydesdorff, 2005).

In general, the majority of scholars using initial-based disambiguation have acknowledged the problem of misidentifying authors (i.e., merging and splitting of identities) when relying on initials for name disambiguation. Some have, however, also argued that disambiguation approaches do not significantly affect research findings. For example, Newman (2001) assumed that the numbers of unique authors identified by first initial and all initial disambiguation correspond to the lower and upper bound of the "true" number of unique authors, respectively. Based on this assumption, he found that most of the statistical properties of coauthorship networks disambiguated by first and all initial methods showed errors or differences of "an order of a few percent." Many scholars cited Newman's approach to justify their use of initial based disambiguation (e.g. Barabasi et al., 2002; Goyal, van der Leij, & Moraga-Gonzalez, 2006; Liben-Nowell & Kleinberg, 2007; Milojević, 2010).

One common problem with initial based disambiguation in coauthorship network studies is that the assumption of the supposedly mild effects of disambiguation errors has not been tested against ground-truth data. An exception here is the work by Milojević (2013), who tested the accuracy of initial based disambiguation on synthetic datasets. However, the accuracy of the simulated data against ground-truth data was not verified. Overall, the identification of biases and errors induced by initial based disambiguation is only possible if ground truth data is available. Since human-disambiguated coauthorship data are extremely rare and only available on a small scale, scholars have been using highly accurate computational solutions as a proxy (Fegley & Torvik, 2013; Strotmann & Zhao, 2012). Even though the most advanced algorithms cannot guarantee perfect disambiguation (Diesner & Carley, 2009), this strategy allows for comparing datasets and results

based on initial based disambiguation and computationally disambiguated datasets. For example, Fegley and Torvik (2013) showed that initial based disambiguation can "dramatically" distort identified collaboration patterns. They compared two coauthorship networks that were generated from the same dataset (9 million names in MEDLINE): one network was disambiguated with advanced algorithms (accuracy of up to 99%) and the other by the all initial method. Through this, they found that all initial disambiguation estimated the number of unique authors as about 2.2 million while their algorithmic disambiguation identified almost 3.2 million unique authors from the same dataset. Strotmann and Zhao (2012) disambiguated names in more than 2.2 million papers from MEDLINE and found that, when disambiguated by the first initial method, about fifty of the top 200 most cited scholars are Asian authors (such as Wang, J.) who are actually merged identities. However, except for the field of biomedicine, where Fegley and Torvik (2013) as well as Strotmann and Zhao (2012) conducted such studies, we have no understanding of the impact of disambiguation strategies and their errors rates on the data and any results computed over the data, nor on any policy implications made based on these results.

The herein presented study is in line with the work by Fegley and Torvik (2013) and Strotmann and Zhao (2012) in that it attempts to estimate the effect of errors that are due to initial based disambiguation by comparing coauthorship networks generated from the same dataset by using different disambiguation methods. Our study differs from prior work in that we consider different domains, namely computer science and information science. There is a rich body of prior work on coauthorship network studies in these fields, and many of the papers used the initial based disambiguation method (e.g. Fiala, 2012; He, Ding, & Ni, 2011; Rorissa & Yuan, 2012). In the following section, we outline the characteristics of the dataset and metrics for measuring network properties used herein.

3. METHODOLOGY

3.1. Data

We used data from DBLP. The DBLP database is a service developed by Dr. Michael Ley at Trier University, Germany (Ley, 2002). Each publication record in DBLP includes at least the author's names as well as title and year of publication. DBLP mainly covers the field of computer science in a broad sense. This includes various key journals from library and information science, such as *Journal of the American Society for Information Science and Technology, Journal of Information Science, Journal of Informetrics, Information Society, Library Trends*, and *Scientometrics*.

DBLP is well known for its high quality citation data. This is partially due to the fact that the DBLP team has dedicated database management efforts to name disambiguation (Ley, 2002, 2009). DBLP uses full names as much as possible, which is believed to alleviate errors with splitting and merging identities. Also, it exploits diverse string matching algorithms as well as coauthorship information to assign names to the presumably correct author identities. For some suspicious cases of split or merged identities, manual inspection is employed. Thus, although DBLP inevitably contains errors, it is internationally respected by computer scientists and information scientists for its accuracy (Franceschet, 2011). Hence, DBLP has been used as a data source for more than 400 scientific studies of name disambiguation, collaboration patterns, and data management (Ley, 2009).

As of March 2014, DBLP contained almost 2.5 million records for journals, proceedings, books and reviews. From these, we selected records of journal papers because the majority of coauthorship studies have focused on journal articles. We retrieved a total of 1,076,577 records of papers published in 1,409 journals, containing 2,812,236 name instances and spanning a period from 1936 to the beginning of 2014. Here, the number of name instances refers to the count of all names in the data regardless of duplicates. This means that, for example, 'Linda Smith' may appear three times in the dataset since she has published three papers, while the count of the unique name is just one. From this dataset, we excluded papers with no author name or authored by a single author. We made this decision since most of the previous coauthorship network studies excluded single authored papers from analysis. This reduction process resulted in 816,643 papers (2,557,898 name instances). Our selected subset contains 75.9 % of all the papers we retrieved and 91.0 % of all name instances

included in all the papers we retrieved. This not only indicates that coauthoring is the norm in these fields, but also further substantiates the need for a precise understanding of the impact of disambiguation on coauthorship networks.

3.2. Generating Coauthorship Networks

In a social network, two agents (nodes) are connected by a line (edge) if they interact with each other, exchange resources, or share an affiliation or activity (Knoke & Yang, 2008). In a coauthorship network, which is a special type of social network, two authors get linked if they coauthor a publication. To test the performance of initial based disambiguation in terms of accuracy, we generated three coauthorship networks from the same dataset. First, by using the raw dataset of 816,643 papers, we constructed a proxy of the ground-truth network. Next, we disambiguated names in the same dataset by using the first and all initial disambiguation methods, and generated a network from each dataset. We excluded the hybrid method from our analysis because it has not been frequently used except in one empirical study by Yoshikane et al. (2009). For the first and all initial disambiguation methods, some preprocessing is required. Name instances in DBLP are represented as given name plus last name (e.g. Linda Smith). This is in contrast to other bibliometrics datasets where an author's name is provided as a last name followed by a given name (e.g. Smith, Linda or Smith, L.). To apply name initial disambiguation, the last name and given name of a name instance should be distinguished from each other. This can mostly be done by locating the last name part in a name string (e.g., Chang in 'Alan Chin-Chen Chang'). The problem is that it is sometimes unclear which name part is a last name or a given name. For example, in some Spanish-speaking countries, the norm is to have two given names and two last names, e.g. "Juan Antonio Holgado Terriza," where the last names are "Holgado Terriza." Again, in most cases, this can be dealt with by locating the last name part that is indicated by a hyphen (e.g., Sangiovanni-Vicentelli in 'Alberto L. Sangiovanni-Vicentelli'). However, some name instances contain no such clue. To deal with such cases, we downloaded records of 57,099 papers published in 99 top journals in computer science (including some journals also categorized into information science) between 2009 and 2013 using the *Journal Citation Report 2013* from the Web of Science (Thompson Reuters). From 185,518 name instances in the last name plus given name format, we extracted 3,892 unique cases of a last names with two or more name parts (e.g., Hernandez del Olmo) and 140 single last name prefixes (e.g., 'das' or 'van de'), and applied this information to detect last names of about 120,000 name instances in our dataset.

3.3. Measurements

We selected the following six metrics since they are commonly used in coauthorship network studies. We used the social network analysis package *Pajek* (de Nooy, Mrvar, & Batagelj, 2011) to compute these metrics on our data.

Productivity: This is measured as the total number of papers per unique author. Merged and split identities directly impact this metric as they can inflate or deflate the number of publications per person.

Number of Coauthors (Degree Centrality): Two scholars are connected if they appear as coauthors on a paper. The degree centrality (short *degree*) of an actor (node) refers to the number of direct connections that he or she has. Here, only the existence of collaboration ties between authors is considered (binary ties), while the number of co-authored papers (ties weighted by frequency) is disregarded. We made this decision to resemble common procedure in coauthorship studies (Barabasi et al., 2002; Moody, 2004; Newman, 2001). In short, the degree of an author represents the number of her unique collaborators.

Density: Density measures the proportion of the number of actual ties over the number of possible ties (excluding self-loops). Network scholars have considered this measure as an indicator of network cohesion (Wasserman & Faust, 1994), although this is controversial (Friedkin, 1981).

Components: A component is a maximal subgraph where any node can reach any other node in one or more steps. Scholars typically look at the size of the largest component and the number of components, which together inform us about the coherence or fragmentation of a network (Newman, 2001).

Shortest Path: The shortest path, also known as the geodesics, between two authors is the minimum number of steps between them. In this study, the average

shortest paths of all authors in each dataset (Brandes, 2008) are reported. Only the lengths of existing paths were averaged.

Clustering Coefficient: The clustering coefficient measures the average fraction of a person's coauthors who have also published together (Newman, 2001). This type of closure can result from three or more people being involved in the same paper or in different papers. We calculate the clustering coefficient as the ratio of the number of triangles over the number of triples (Fegley & Torvik, 2013). Here, a triangle is a set of three authors who are connected to one another (i.e. via three ties), while a triple is a set of three authors who are held together by exactly two ties.

A*ssortativity*: This measures the extent to which authors collaborate with others who are similar to them in terms of degree. In this study, assortativity is calculated as "the Pearson correlation coefficient of the degrees at either ends of an edge" between any two authors (Newman, 2002).

4. RESULTS

4.1. Number of Unique Authors

The number of identified unique authors from DBLP is shown in Table 2. The last column in this table shows the reduction in the number of individuals – i.e., the effect of collapsing multiple truly distinct authors into clusters of people who happen to share the same name – compared against the DBLP data. Here, DBLP serves as a proxy for ground truth. Overall, the two considered initial based disambiguation methods underestimate the number of unique authors by 30% (all initial method) to 43% (first initial method). This indicates that initial based disambiguation will suggest smaller scholarly communities in computer and information science than there really are.

Table 2. Number of Unique Authors per Method

	Number of Unique Authors	Change (%)
DBLP	775,854	-
All-Initial Method	545,072	-29.75
First-Initial Method	440,981	-43.16

Moreover, our results suggest that the number of unique individuals is greater than the upper bound for unique authors when the upper bound is computed as the largest number of unique authors estimated by the all initial method. This finding contradicts prior work by some researchers, e.g. Newman (2001), but confirms results by others, e.g. Fegley and Torvik (2013). The latter reported that the total number of unique authors as identified by their algorithmic disambiguation (3.17×10^6) exceeded the upper bound generated by the all initial method (2.18×10^6, -31.23%). Also, the ratio of decrease identified by Fegley and Torvik (2013) (-31.23%) is similar to the ratio we found (-29.75%).

4.2. Distributions of Productivity and Number of Coauthors

The underestimation in the number of unique authors indicates that initial based disambiguation merges author identities that should actually be split apart. To illustrate the effect of merged (or split) author identities on statistical properties, the distributions of productivity and number of coauthors (i.e. degree) are shown in Figures 1 and 2. For each cumulative log-log plot, we show the distributions from all three datasets we used: DBLP (red triangles), all initial method (black crosses), and first initial method (blue circles).

The curves of the original DBLP dataset show more downwards curvature compared to those disambiguated by initials. This finding means that, for a given value (x) of productivity or degree, the proportion of authors who have the given value ($X = x$) or a value above the given value ($X > x$) increases by initial based disambiguation for the majority of x values. In other words, the blue circle and black cross curves positioned above the red triangle curves indicate that the productivity and degree distributions are distorted: initial based disambiguation methods create merged identities, which leads to an inflation of values for productivity and degree. This pushes the curves both upwards and to the right. Moreover, the curves from both initial based disambiguation methods seem to have a lower, straighter slope than those from the DBLP data, which might be fit by a power-law distribution or Lotka's Law (Barabasi et al., 2002; Huber, 2002; Newman, 2001). Fegley and Torvik (2013, Figure 8) came to the same conclusion: degree distribution from the algorithmic disambiguation shows much more curvature than the

one based on all initial based disambiguation. Overall, these findings indicate that the average productivity and number of collaborators can be distorted by initial based disambiguation such that scholars in a target dataset are portrayed as more productive and collaborative than they really are.

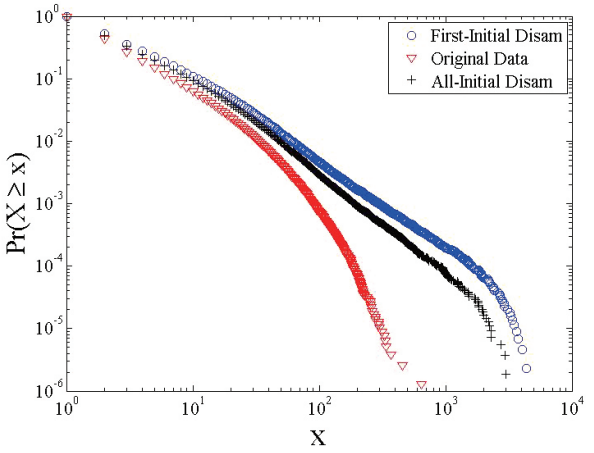

Fig. 1 Cumulative log-log plot of productivity distribution

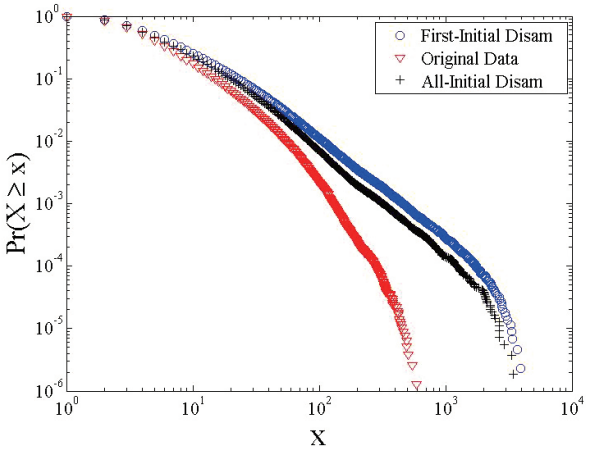

Fig. 2 Cumulative log-log plot of degree distribution

4.3. Statistical Properties from Network Metrics

We report additional statistical properties of the three coauthorship networks in Table 3. First, disambiguation has a much smaller impact on the number of ties than the number of nodes. More precisely, the first initial method reduced the number of ties by 6.03%, and the all initial method by 2.90%. This suggests that merged author nodes typically have distinct sets of coauthors. If two merged author identities have coauthors that are also merged due to name ambiguity, then the ties between each author and coauthor would also be merged, leading to the decrease of ties.

Using initial based disambiguation leads to method-induced increases in network density, average productivity, degree, and the size of the largest component. These increases are expected as they are logical consequences of merging actually distinct people into collective persona. This procedure causes a higher number of publications and collaborators per node and a reduction in the total number of unique authors. When using the all initial method and first initial method, respectively, density doubled and tripled, degree increased by over a third to two thirds, and productivity went up by 43% to 75% – all due to data pre-processing decisions instead of any change in underlying social behavior.

At the same time, we observe inaccurate decreases in the average shortest path length, clustering coefficients (a.k.a. transitivity), assortativity, and the number of components due to using initial based disambiguation. The clustering coefficient is measured as the fraction of triangles (3 nodes with 3 ties between them) over the number of triples connected with exactly two ties. When using initial based disambiguation, the merging of author identities leads to a stronger increase in the number of connected triples (denominator) than the number of triangles (numerator), which again is an expected mathematical consequence. Overall, networks generated with initial based disambiguation are more coherent and interconnected than the underlying true network is, and individual authors appear to be more productive and more strongly embedded than they actually are.

Our findings are consistent with those of Fegley and Torvik (2013) and Velden, Haque, and Lagoze (2011). We also find that the numerical differences between the coauthorship network generated from a reasonable proxy for ground truth versus the networks generated from the same data after pre-processing it with initial based disambiguation methods exceed "the order of a few percent" (Newman 2001).

Table 3. Overview of Statistical Properties of Networks per Method

	DBLP	All-Initial Method	First-Initial Method
No. of Ties	2,660,700	2,583,615	2,500,186
Density	8.84E-06	1.73E-05	2.57E-05
Avg. Productivity (SD)	3.30 (7.33)	4.69 (21.23)	5.80 (34.55)
Avg. Degree (SD)	6.86 (11.53)	9.48 (30.33)	11.34 (41.22)
No. of Components (Ratio of Largest Component Size)	38,008 (84.66%)	16,206 (91.82%)	9,012 (94.68%)
Avg. Shortest Path (Path of Most Distant Nodes)	6.56 (25)	5.18 (18)	4.74 (15)
Clustering Coefficient	0.274	0.105	0.096
Assortativity	0.170	0.106	0.094

5. CONCLUSION AND DISCUSSION

This paper attempts to estimate the impact of initial based disambiguation on coauthorship networks. We disambiguated the DBLP dataset of 0.8 million journal papers by first and all initial methods and compared typically used statistical properties of the resulting networks against each other and to a proxy for ground truth data. We conclude that initial based disambiguation can lead to distorted findings and inaccurate representations of scientific collaborations. When using initial based disambiguation, authors become embedded in ties for which there is no empirical support, which leads to increases in people's spheres of influence and diversity of involvement. As more authors get integrated into larger components, some of them seem to serve as bridges connecting previously disjoint (groups of) authors and to provide shortcuts for connecting people. Overall, initial based disambiguation suggests more cohesive networks and more prolific and integrated authors than actually exist. These are wrongfully induced consequences of data pre-processing choices with potentially strong implications for our understanding and modeling of the patterns and dynamics of scientific collation.

For a selected set of network properties, we showed an overall decrease in network analytical values due to initial based disambiguation. More specifically, this applies to the number of unique authors and collaboration ties, average shortest path length, clustering coefficient, assortativity, and the number of components. Other measures increased: network density, author's productivity and degree, and the size of the largest component. In summary, these effects imply that initial based disambiguation produces coauthorship networks that are smaller and less fragmented than the true underlying network is, and represents those networks as ones where people can reach each other more efficiently, are more productive, and have a larger and more diverse set of collaborators.

This study is not without limitations. The findings involve some domain-specificity as our data originate mainly from a specific dataset from computer science and information science. Additional studies on other fields are needed to generalize these conclusions. Second, it is unknown how the distortive effects of initial based disambiguation impact smaller datasets. As name ambiguity increases with the size of the dataset (Fegley & Torvik, 2013), one may expect that in a small scale coauthorship network initial based disambiguation is less detrimental than when applied to larger datasets. As far as we know, our study used the second largest coauthorship network for measuring the impact of initial based disambiguation following Fegley and Torvik (2013). Third, this study lacks detailed explanation of what factors affect the errors of authorship identification by initial based disambiguation. For example, scholars have suggested that

Asian names, especially Chinese, Japanese, and Korean names, contribute more to name ambiguity in authorship identification as they are known to share common last names (Strotmann & Zhao, 2012; Torvik & Smalheiser, 2009). In our study, the majority of the top 100 names that appear frequently when initial based disambiguation applied were Asian names such as Kim, Lee, Zhang, and Wang. The extent to which these ambiguous names cause authorship misidentification may provide a deeper understanding of disambiguation issues. The outlined limitations are topics of our future research.

The main takeaway from this study is that initial based disambiguation can underestimate the number of authors and connections mainly through merging, and, therefore, can distort macroscopic views of the patterns and evolution of collaboration. This implies that coauthorship network research, especially when done based on large scale data, should pay more attention to name ambiguity, and any findings should be treated with caution when names are not properly disambiguated.

ACKNOWLEDGEMENTS

This work is supported by KISTI (Korea Institute of Science and Technology Information), grant P14033, and the FORD Foundation, grant 0145-0558. We would like to thank Vetle Torvik and Andrew Higgins for their comments on this manuscript.

REFERENCES

Barabasi, A. L., Jeong, H., Neda, Z., Ravasz, E., Schubert, A., & Vicsek, T. (2002). Evolution of the social network of scientific collaborations. *Physica a-Statistical Mechanics and Its Applications, 311*(3-4), 590-614. doi: 10.1016/s0378-4371(02)00736-7

Bettencourt, L. M. A., Lobo, J., & Strumsky, D. (2007). Invention in the city: Increasing returns to patenting as a scaling function of metropolitan size. *Research Policy, 36*(1), 107-120. doi: 10.1016/j.respol.2006.09.026

Brandes, U. (2008). On variants of shortest-path betweenness centrality and their generic computation. *Social Networks, 30*(2), 136-145. doi: http://dx.doi.org/10.1016/j.socnet.2007.11.001

Braun, T., Glanzel, W., & Schubert, A. (2001). Publication and cooperation patterns of the authors of neuroscience journals. *Scientometrics, 51*(3), 499-510. doi: 10.1023/a:1019643002560

de Nooy, W., Mrvar, A., & Batagelj, V. (2011). *Exploratory social network analysis with Pajek: Cambridge University Press.*

Diesner, J., & Carley, K. M. (2009). He says, she says, pat says, Tricia says: how much reference resolution matters for entity extraction, relation extraction, and social network analysis. Paper presented at *the Proceedings of the Second IEEE international conference on Computational intelligence for security and defense applications,* Ottawa, Ontario, Canada.

Fegley, B. D., & Torvik, V. I. (2013). Has Large-Scale Named-Entity Network Analysis Been Resting on a Flawed Assumption? *Plos One, 8*(7). doi: 10.1371/journal.pone.0070299

Fiala, D. (2012). Time-aware PageRank for bibliographic networks. *Journal of Informetrics, 6*(3), 370-388. doi: 10.1016/j.joi.2012.02.002

Franceschet, M. (2011). Collaboration in Computer Science: A Network Science Approach. *Journal of the American Society for Information Science and Technology, 62*(10), 1992-2012. doi: 10.1002/asi.21614

Friedkin, N. E. (1981). The Development of Structure in Random Networks: An Analysis of the Effects of Increasing Network Density on Five Measures of Structure. *Social Networks, 3*(1), 41-52.

Goyal, S., van der Leij, M. J., & Moraga-Gonzalez, J. L. (2006). Economics: An emerging small world. *Journal of Political Economy, 114*(2), 403-412. doi: 10.1086/500990

He, B., Ding, Y., & Ni, C. (2011). Mining Enriched Contextual Information of Scientific Collaboration: A Meso Perspective. *Journal of the American Society for Information Science and Technology, 62*(5), 831-845. doi: 10.1002/asi.21510

Huber, J. C. (2002). A new model that generates Lotka's Law. *Journal of the American Society for Information Science and Technology, 53*(3), 209-219. doi: 10.1002/asi.10025

Knoke, D., & Yang, S. (2008). *Social network analysis.* Los Angeles, CA: Sage Publications.

Lariviere, V., Sugimoto, C. R., & Cronin, B. (2012). A

bibliometric chronicling of library and information science's first hundred years. *Journal of the American Society for Information Science and Technology, 63*(5), 997-1016. doi: 10.1002/asi.22645

Lee, D., Goh, K. I., Kahng, B., & Kim, D. (2010). Complete trails of coauthorship network evolution. *Physical Review E, 82*(2). doi: 10.1103/PhysRevE.82.026112

Ley, M. (2002). The DBLP Computer Science Bibliography: Evolution, Research Issues, Perspectives. In A. F. Laender & A. Oliveira (Eds.), *String Processing and Information Retrieval* (Vol. 2476, pp. 1-10): Springer Berlin Heidelberg.

Ley, M. (2009). DBLP: some lessons learned. Proc. *VLDB Endow., 2*(2), 1493-1500.

Leydesdorff, L., & Sun, Y. (2009). National and International Dimensions of the Triple Helix in Japan: University-Industry-Government Versus International Coauthorship Relations. *Journal of the American Society for Information Science and Technology, 60*(4), 778-788. doi: 10.1002/asi.20997

Liben-Nowell, D., & Kleinberg, J. (2007). The link-prediction problem for social networks. *Journal of the American Society for Information Science and Technology, 58*(7), 1019-1031. doi: 10.1002/asi.20591

Milojević, S. (2010). Modes of Collaboration in Modern Science: Beyond Power Laws and Preferential Attachment. *Journal of the American Society for Information Science and Technology, 61*(7), 1410-1423. doi: 10.1002/asi.21331

Milojević, S. (2013). Accuracy of simple, initials-based methods for author name disambiguation. *Journal of Informetrics, 7*(4), 767-773. doi: http://dx.doi.org/10.1016/j.joi.2013.06.006

Moody, J. (2004). The structure of a social science collaboration network: Disciplinary cohesion from 1963 to 1999. *American Sociological Review, 69*(2), 213-238.

Newman, M. E. J. (2001). The structure of scientific collaboration networks. *Proceedings of the National Academy of Sciences of the United States of America, 98*(2), 404-409. doi: 10.1073/pnas.021544898

Newman, M. E. J. (2002). Assortative mixing in networks. *Physical Review Letters, 89*(20), 208701.

Rorissa, A., & Yuan, X. J. (2012). Visualizing and mapping the intellectual structure of information retrieval. *Information Processing & Management, 48*(1), 120-135. doi: 10.1016/j.ipm.2011.03.004

Smalheiser, N. R., & Torvik, V. I. (2009). Author Name Disambiguation. *Annual Review of Information Science and Technology, 43*, 287-313.

Strotmann, A., & Zhao, D. Z. (2012). Author name disambiguation: What difference does it make in author-based citation analysis? *Journal of the American Society for Information Science and Technology, 63*(9), 1820-1833. doi: Doi 10.1002/Asi.22695

Torvik, V. I., & Smalheiser, N. R. (2009). Author Name Disambiguation in MEDLINE. *Acm Transactions on Knowledge Discovery from Data, 3*(3). doi: Doi 10.1145/1552303.1552304

Torvik, V. I., Weeber, M., Swanson, D. R., & Smalheiser, N. R. (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. doi: Doi 10.1002/Asi/20105

Treeratpituk, P., & Giles, C. L. (2009). Disambiguating Authors in Academic Publications using Random Forests. *Paper presented at the Jcdl 09: Proceedings of the 2009 Acm/Ieee Joint Conference on Digital Libraries.*

Velden, Haque, A., & Lagoze, C. (2011). Resolving author name homonymy to improve resolution of structures in co-author networks. *Paper presented at the Proceedings of the 11th annual international ACM/IEEE joint conference on Digital libraries.*

Wagner, C. S., & Leydesdorff, L. (2005). Network structure, self-organization, and the growth of international collaboration in science. *Research Policy, 34*(10), 1608-1618. doi: http://dx.doi.org/10.1016/j.respol.2005.08.002

Wasserman, S., & Faust, K. (1994). *Social network analysis: Methods and applications.* New York, NY: Cambridge University Press.

Yoshikane, F., Nozawa, T., Shibui, S., & Suzuki, T. (2009). An analysis of the connection between researchers' productivity and their co-authors' past attributions, including the importance in collaboration networks. *Scientometrics, 79*(2), 435-449. doi: 10.1007/s11192-008-0429-8