Why Name Ambiguity Resolution Matters for Scholarly Big Data Research

Jinseok Kim & Jana Diesner
Graduate School of Library and Information Science
University of Illinois at Urbana-Champaign
Urbana, USA
{jkim362, jdiesner}@illinois.edu

Amirhossein Aleyasen
Department of Computer Science
University of Illinois at Urbana-Champaign
Urbana, USA
Aleyase2@illinois.edu

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

Hwan-Min Kim
Department of Overseas Information
Korea Institute of Science and Technology Information
Daejeon, Korea
mrkim@kisti.re.kr

Abstract— **This paper illustrates how data pre-processing choices about author name disambiguation can affect research findings about scholarly networks and hypotheses about underlying social mechanisms. We have analyzed three big scholarly datasets that were disambiguated algorithmically and via two common initial-based disambiguation methods; namely first-initial and all-initials disambiguation. The comparison of resulting bibliometric and network properties revealed that initial-disambiguation bears the prevalent risks of incorrectly merging author identities, underestimating the number of unique authors and inflating the average productivity and number of collaborators per author. The gaps between outcomes of name ambiguity resolution methods range from -4.23% to -87.36% per dataset for the number of unique authors, from 3.75% to 691.20% for average productivity, and from 5.06% to 285.28% for degree centrality for initial based methods compared to algorithmic disambiguation. This calls for special attention to data pre-processing choices in scholarly big data research.**

Keywords— collaboration; network analysis; disambiguation; bibliometrics

I. Introduction

Since Mark Newman leveraged computational techniques to study large-scale coauthorship networks, many scholars have contributed to gaining a birds-eye view of the patterns and dynamics of scientific collaboration in a wide range of fields and publication venues [1]. For example, in a coauthorship study of more than 210,000 articles in neuroscience and 70,000 articles in mathematics, scholars were found to select collaborators who have already many collaborators [2]. The underlying preferential attachment mechanism has been shown to lead to scale-free networks, where the distribution of the number of coauthors per author has a slope that can be described with a power law [2-4]. It has also been shown that scholars can reach one another within a few steps of collaboration relationship, which confirms the small world property of scientific collaboration [1, 5].

Such macroscopic views of the structure of scientific collaboration have been usually obtained from analyses of large-scale bibliometric data. For instance, more than 2 million MEDLINE records published between 1995 and 1999 were analyzed by [4]. Other prominent examples include studies of coauthorship networks from specific fields, including mathematics (1.6 million papers, 1940-1999) [6], computer science (1.2 million papers, 1936-2008) [1], physics (458,799 papers, 1893-2009) [7], sociology (281,090 papers, 1963-1999) [8], and specific geographical regions, e.g. Turkey (237,409 papers, 1980-2010) [9].

One common challenge for scholars dealing with these instances of big data has been the problem of name ambiguity. As pointed out as early as in the late 1960s [10], the tradition of representing author names in the format of full surname followed by first or middle name initials have been a source of “ambiguity and confusion”. For example, this ambiguity resolution convention for bibliometric records can cause two truly distinct authors who happen to share the same surname and first name initial, e.g., ‘Blake, Catherine’ and ‘Blake, Cooper’, to be merged into one identity, in this case ‘Blake, C.’. This effect can have repercussions on the networks structure, conclusions we draw, and theories we build about the patterns and evolution of co-authorship networks, including the identification of main key players and sub-communities in a field.

Several strategies for dealing with this issue inherent to studying big bibliometric data have been established: For one, people have relied on disambiguation conducted by data providers, such as DBLP, and/ or devised (their own) heuristics, rules, methods and algorithms of solving name ambiguity in their datasets. Some scholars employed advanced disambiguation algorithm [11, 12]. Many have, however, relied on a simple heuristic, which can be considered common practice in bibliometrics: namely, initial-based name

This work is supported by KISTI (Korea Institute of Science and Technology Information), grant P14033.

disambiguation. With one version of this approach, author names overlapping in surname and first name initial are considered to refer to the same person ('first-initial method' hereafter, e.g., [13, 14]). A more common strategy is to consider all initials of first and middle names: author names with the same surname as well as same all given name initials are seen to represent the same author ('all-initials method' hereafter, e.g., [3, 15]).

Overall, using initial-based name disambiguation for pre-processing big data bears the risk of misidentifying authors by incorrectly merging or splitting author names. This is a natural yet problematic side effect of the author name ambiguity resolution. For example, 'Blake, C. L.' would be the same person as 'Blake, C. S.' according to the first-initial method. This is a 'merging' error. Also, while 'Blake, C.' and 'Blake, C. L.' would be different authors according to the all-initials method, they would be the same person as 'Blake, C. L.' in cases where the middle name initial is omitted. This is a 'splitting' error. The merging and splitting errors have been well acknowledged by scholars who have used name initials for identifying authors in bibliometric data. They have, however, justified their initial based disambiguation with the following explanations. First, even the most advanced disambiguation algorithms do not guarantee perfect disambiguation. This is further pronounced by the common effect that publication information is often imperfect due to missing or inconsistent recording [16-18].

Second, it has been assumed that such misidentification errors do not affect research findings too drastically in terms of both macroscopic and microscopic views [2, 14, 19]. The problem with this justification is that the argument is often based on (1) an assumption that numbers of authors identified by all-initial methods and first-initial method respectively represent the upper and lower bound of the 'true' number of authors and (2) prior empirical work showing that statistical properties of coauthorship networks generated from these two sets of authors show errors of 'an order of a few percent' [2, 3, 13, 14].

Recently, some scholars have tested this assumption by comparing the statistical properties of networks generated from bibliometric data that were disambiguated by algorithmic solutions versus by initial-based disambiguation [12, 19-21]. These studies have shown that initial-based disambiguation can misrepresent network properties to a non-negligible extent. One example is that the number of unique authors (3.2 million) identified by algorithmic disambiguation was twice the number (1.56 million) identified by first-initial method in MEDLINE data [20]. Another study, however, found that, based on simulated authorship data across five fields, initial-based disambiguation to be 'quite accurate' at identifying authors, and that first-initial method is superior to all-initials method [19]. Similarly, another study argued that the influence of name ambiguity on research finding is limited, e.g., if only authors who appear as the last author in each paper are considered [12].

In this way, the effect of pre-processing method on research output in bibliometric data analysis is still controversial and not sufficiently understood. Considering the fact that initial-based disambiguation is a dominant method in bibliometric studies [12, 19, 21], testing the effect of name ambiguity on statistical characteristics of data is of great importance. This paper aims to enrich this discussion by illustrating how the choice of data pre-processing methods affects our understanding of productivity and collaboration pattern in selected bibliometric dataets. The contribution with this study lies in providing a better understanding of the impact of name disambiguation methods in two scientific fields (computer science and physics), which have been a frequent subject in bibliometric research, and domestic-level collaboration among Korean scholars.

II. Data

A. DBLP

The Digital Bibliography & Library Project (DBLP) database provides publication records of conference and journal papers in computer science. DBLP has been used by many computer and information scientists to study collaboration structure, test name disambiguation algorithms, or data management (e.g., [1, 21-23]). Such a wide use of DBLP is partially due to its data quality: author names are disambiguated both algorithmically and manually by the DBLP management team [24, 25]. For our study, we retrieved publication records of 315,828 journal papers between 2005 and 2009 (5 year window). As DBLP does not distinguish surname and given name parts of each name instance, we followed the method described in [21] to format each name instance into a surname followed by a given name.

B. APS

The second dataset was obtained from the American Physical Society (APS), who provides publication records of the Physical Review journals; a family of journals published by the APS covering all subfields of physics. The dataset has been used by numerous scholars to map scientific collaboration in physics and to test theories or algorithms related to the study of complex dynamic networks (e.g., [7, 15, 26, 27]). As author names in the raw APS data are not disambiguated, we applied the same algorithm described in [7] for name disambiguation, where names were clustered based on the similarity of name string, coauthor name, affiliation, and venue information. From the disambiguated dataset, we selected publication records of 90,784 papers spanning from 2005-2009.

C. KISTI

Nation-level publication data for scholars in Korea was obtained from the National Digital Science Library (NDSL). This dataset was built and managed by the Korea Institute of Science and Technology Information (KISTI), a government-funded organization gathering, analyzing, and distributing scientific publication information abroad as well as nationwide. Asian names such as Chinese and Korean names, are known to be more difficult to disambiguate than Western names, which is mainly due to the large ratio of shared surnames and given names [12, 28]. KISTI went through a two-step process to disambiguate author names. First, author name instances were clustered based on algorithms using feature vectors like full name string, affiliation, coauthor name, title, and publication outlet. Then, the suspicious clustering results were manually verified by human specialists. For this study, we collected publication records of 161,569 journal papers published in Korea from 2005 to 2009. The majority of author names in the KISTI data are also recorded in English. Some names in Korean were automatically changed into English.

III. Measurements

A. Network Generation

Each data (we refer them to DBLP, APS, and KSITI hereafter) was furthermore disambiguated by first- and all-initials methods. In total, we generated three coauthorship networks from for each of the three datasets: (1) the first ones by algorithmic disambiguation performed by DBLP management team (DBLP), Newman's team (APS), and KISTI team (KISTI), respectively, (2) the second ones from the DBLP, APS, and KISTI data disambiguated by first-initial method, (3) while the third ones were disambiguated by all-initials method.

B. Measures

We use the following state of the art network metrics for assessing the structure of the generated network datasets.

- Number of Unique Authors: This represents the number of unique author identities identified by algorithmic, first-initial, and all-initials disambiguation methods.
- Productivity: An author's productivity is the total number of papers per unique author. This is calculated as the name frequency of an author who is uniquely identified per disambiguation method.
- Degree (Number of Coauthors): Two authors are connected by a coauthor relationship if they appear as coauthors on the same paper. The degree centrality (in short, degree) of an actor is the number of direct connections that he or she has. Here, only the existence of collaboration ties between authors is considered following the convention of previous coauthorship network studies [2, 4, 8].

IV. Analysis

A. Number of Unique authors

Table 1 shows the numbers of unique authors in each dataset identified by the different disambiguation methods: algorithmic, first-initial, and all-initials methods. Here, algorithmically disambiguated data serves as baseline data against which we compare the initial-based disambiguation approaches. Overall, initial-based disambiguation underestimates the numbers of unique authors by 4.23% (APS, all-initials method) to 87.36% (KISTI, first-initial method). This implies that if bibliometric data is pre-processed by initial-based disambiguation, researchers may discover smaller scholarly communities than there actually are. In addition, the overall underestimation implies that merging is more prevalent than splitting. This argument is, of course, based on the assumption that algorithmically disambiguated data approximate ground-truth data. Interestingly, KISTI seems to be an extreme case showing how distortive the effect of name disambiguation methods can be. This might be due to the fact that almost half of the Korean people share the same three surnames, i.e., Kim, Lee, and Park [29].

TABLE I. Number of Unique Authors per Disambiguation Method (ratio of change against algorithmic data in parentheses)

Data	Algorithmic	First-initial	All-initials
DBLP	354,764	227,661 (−35.83%)	271,179 (−23.56%)
APS	101,455	83,914 (−17.29%)	97,168 (−4.23%)
KISTI	164,351	20,778 (−87.36%)	41,425 (−74.79%)

Another noticeable finding is that the upper-bound provided by all-initials method is smaller than that of the algorithmically disambiguated data (see 'All-initials' column in Table 1). This confirms the findings of some prior works (e.g., [20]), but contradicts other work based on the assumption that the all-initials method can provide the upper-bound for the number of unique authors (e.g., [4]).

B. Productivity

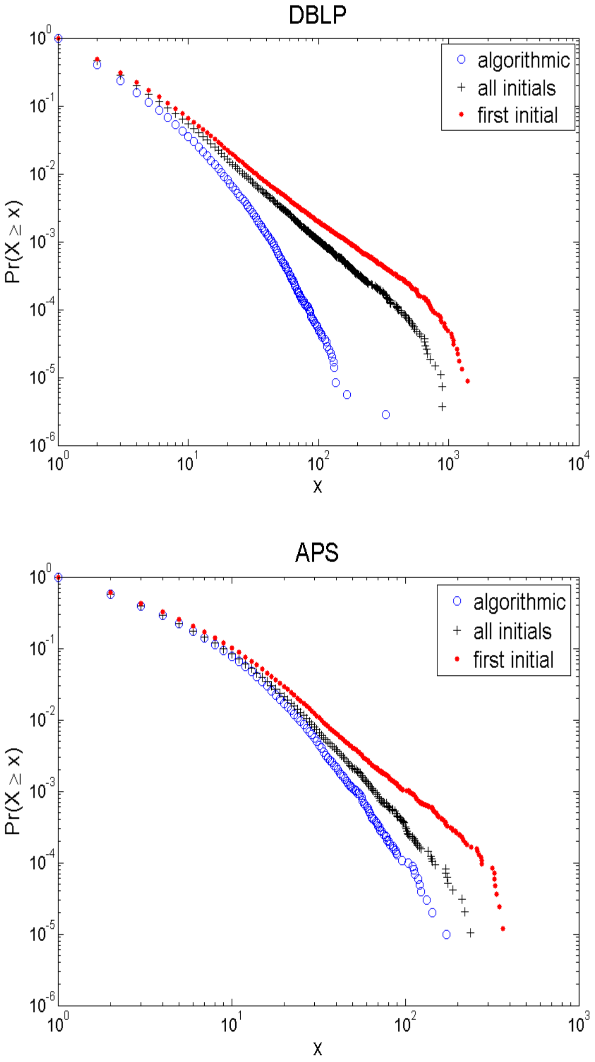
Table 2 shows the average productivity, i.e., the number of papers, per author in each dataset. Across all datasets, initial-based disambiguation overestimates the average productivity by 4.46% (APS, all-initials method) to 691.20% (KISTI, first-initial method). This phenomenon is expected: as initial-based disambiguation mostly merges different author identities into a single author, publications of merged authors will be assigned to a single person.

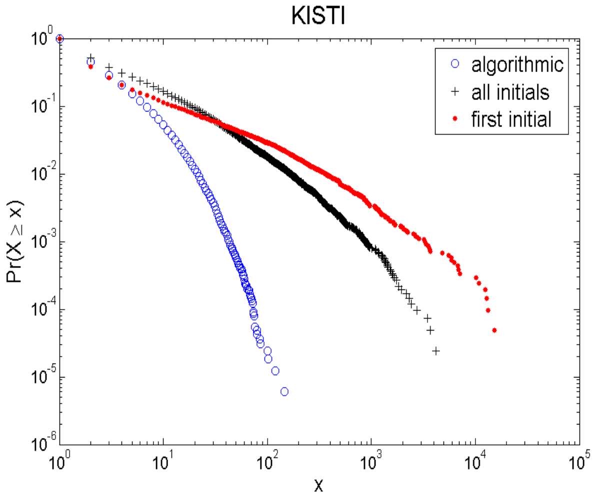
TABLE II. Average Productivity of Authors per Disambiguation Methods (ratio of change against algorithmic data in parentheses)

Data	Algorithmic	First-initial	All-initials
DBLP	2.44	3.81 (+56.15%)	3.20 (+31.15%)
APS	3.59	4.34 (+20.89%)	3.75 (+4.46%)
KISTI	2.84	22.47 (+691.20%)	11.27 (+296.83%)

Figure 1 shows the distributions of productivity in a cumulative log-log plot. In each subfigure, the distribution of algorithmically disambiguated data is shown in blue circles, all-initials method in black crosses, and first-initial in red dots. One common feature of plots across all subfigures is that the distribution of algorithmic disambiguation data shows more negative curvature when compared to first- and all-initials disambiguated versions of data. The interpretation of this finding is that, for a given value (x) of productivity, the proportion of authors who have the given value (X = x) or values higher than the value (X > x) is increased by initial-based disambiguation. The reason for this phenomenon is the same shown above in Table 2: initial-based disambiguation mostly consolidates different author identities into one, which inflates the number of papers per author. In terms of graphical movement, this merging effect pushes the distribution curves both upwards and to the right. Overall, the distribution plots of first-initial method in each data shows more upward and to-the-right movement than all-initials method, which indicates that the merging errors in first-initial disambiguation occur more severely than in all-initials method.

Fig. 1. Cumulative Log-Log Plot of Productivity Distributions in Each Data.





C. Degree (Number of Collaborators)

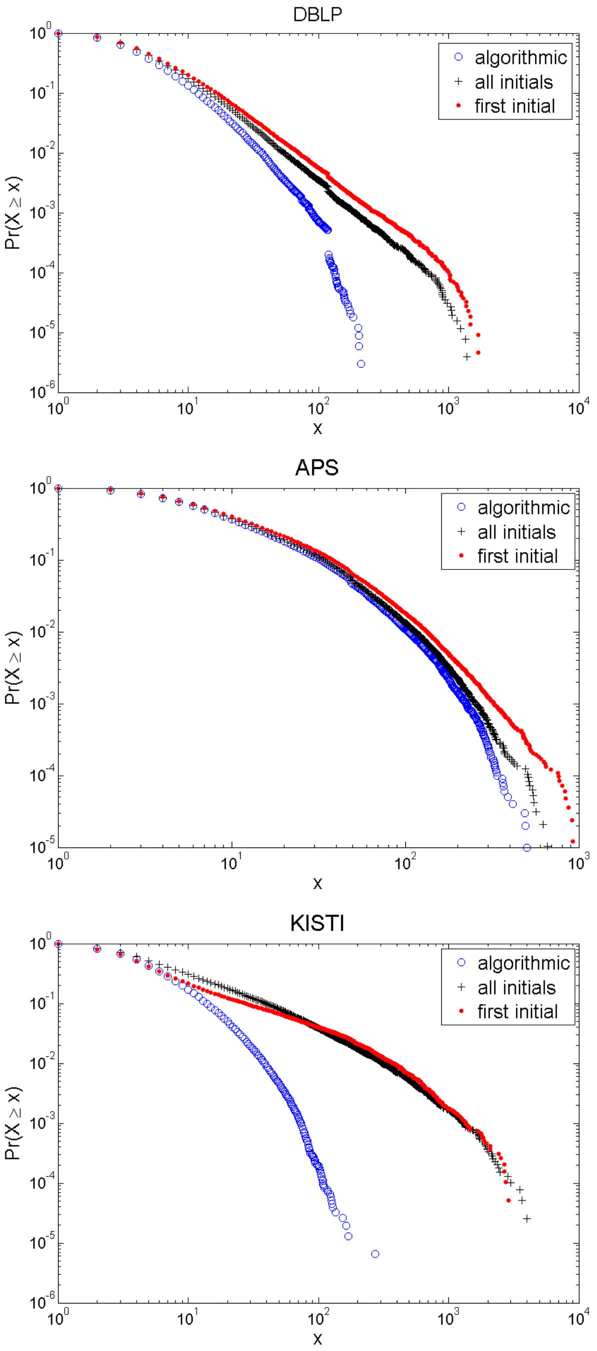
Table 3 shows the average degree (i.e., the number of unique collaborators) of authors in each data. The average degree of authors in algorithmically disambiguated datasets decreased by 5.06% (all-initials in APS) up to 255.28% (first-initial in KISTI). The overall interpretation is the same as average productivity detailed above for Table 2. As initial-based disambiguation merges different author identities into a single person, collaborators of those merged authors get connected to an author who they did not actually collaborate with.

TABLE III. AVERAGE DEGREE OF AUTHORS PER DISAMBIGUATION METHODS (RATIO OF CHANGE AGAINST ALGORITHMIC DATA IN PARENTHESES)

Data	**Algorithmic**	**First-initial**	**All-initials**
DBLP	5.52	8.26 (+49.64%)	7.12 (+28.99%)
APS	13.44	15.92 (+18.45%)	14.12 (+5.06%)
KISTI	6.15	20.75 (+237.40%)	21.85 (+255.28%)

This mechanism also produces the upward and to-the-right moved curves in Figure 2 for the cumulative log-log plot of average degree in each dataset. Authors are shown to have more collaborators when the data are disambiguated by initial-based methods. The gaps between plots are pronounced in the tail area. This indicates that the merging effect can be more easily seen in the top author groups. For example, in DBLP, the most collaborative author (highest degree) has 212 coauthors in the algorithmically disambiguated data, 1,684 in first-initial data and 1,390 in all-initials data. Interestingly, in DBLP, for the majority range of values and in APS for certain range of values, the first-and all-initials methods produce straighter lines than algorithmic disambiguation does, which can be fit by the so-called power distribution with certain "cut-offs".

Fig. 2. Cumulative Log-Log Plot of Degree Distributions in Each Data



V. CONCLUSION AND DISCUSSION

This paper reports on how choices for data pre-processing methods, namely name disambiguation in bibliometric data, can affect our understanding of the fundamental structure of scholarly communities, and lead to biased assumptions about underlying generative processes. We have shown that in contrast to algorithmically disambiguated data, which we consider as a baseline, the two initial-based disambiguation methods tested herein, i.e., first-initial and all-initial methods, incorrectly merge a substantial amount of different author identities. Due to this effect, the size of the scholarly community is underestimated, while the average productivity, connectivity and embeddedness of authors are artificially inflated. These methodologically induced biases can lead to finding incorrect pattern in the data, e.g., a power-law distribution of node degree.

We are not refuting any previous studies where initial-based disambiguation has been used. Also, the selection of datasets and time period (5 year window) entails idiosyncrasies that might not generalize to other data. As we have shown, the magnitude of the effect heavily depends on the data. In other words, each scholarly dataset may have distinct levels or features of name ambiguity, which may reduce the effect of name disambiguation choice to a negligible extent. For example, in APS, all-initials method was found to approximate algorithmic disambiguation with a few percent of errors, while in KISTI, the difference was tremendous. In case like the latter, making an effort to conduct accurate ambiguity resolution will make a big difference in terms of findings and conclusions. In addition, the disambiguation results from three different methods should be compared to ground-truth data to measure the extent of author misidentification, which can provide scholars a guideline on conditions about when to use each of methods safely or avoid one.

The main take away from this study is that research findings from scholarly data should be consumed and assessed with caution when details on the employed name disambiguation strategy are missing. To look at this from a proactive side, our findings call for well-designed, rigorous studies that include the identification of the possible distortive effect of name ambiguity on knowledge discovery in scholarly data. Especially, bibliometric studies of large-scale scholarly data or data on fields where authors with ambiguous names are dominant (e.g., nanoscience) are advised to pay more attention to the importance of data provenance management. This argument is further supported by the fact that name ambiguity increases with the number of author names [20] and certain names such as Asian or Hispanic names are more relevant with respect to ambiguity than others [12, 21].

ACKNOWLEDGMENT

We thank Brian Karrer (Facebook), Travis Martin (Univ. of Michigan), Brian Ball (Dotomi Inc.), and Mark Newman (Univ. of Michigan) for helping us to disambiguate author names in the APS dataset.

REFERENCES

[1] Franceschet, M., Collaboration in Computer Science: A Network Science Approach. Journal of the American Society for Information Science and Technology, 2011. 62(10): p. 1992-2012.

[2] Barabási, A.L., et al., Evolution of the social network of scientific collaborations. Physica a-Statistical Mechanics and Its Applications, 2002. 311(3-4): p. 590-614.

[3] Milojević, S., Modes of Collaboration in Modern Science: Beyond Power Laws and Preferential Attachment. Journal of the American Society for Information Science and Technology, 2010. 61(7): p. 1410-1423.

[4] Newman, M.E.J., The structure of scientific collaboration networks. Proceedings of the National Academy of Sciences of the United States of America, 2001. 98(2): p. 404-409.

[5] Newman, M.E.J., Models of the small world. Journal of Statistical Physics, 2000. 101(3-4): p. 819-841.

[6] Grossman, J.W., Patterns of collaboration in mathematical research. SIAM News, 2002. 35(9): p. 8-9.

[7] Martin, T., et al., Coauthorship and citation patterns in the Physical Review. Physical Review E, 2013. 88(1).

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

[9] Çavuşoğlu, A. and İ. Türker, Scientific collaboration network of Turkey. Chaos, Solitons & Fractals, 2013. 57: p. 9-18.

[10] Garfield, E., British quest for uniqueness versus American egocentrism. Nature, 1969. 223(5207): p. 763.

[11] Kang, In-Su et al., On co-authorship for author disambiguation. Information Processing & Management, 2009. 45(1): p. 84-97.

[12] Strotmann, A. and D. Zhao, Author name disambiguation: What difference does it make in author-based citation analysis? Journal of the American Society for Information Science and Technology, 2012. 63(9): p. 1820-1833.

[13] Liben-Nowell, D. and J. Kleinberg, The link-prediction problem for social networks. Journal of the American Society for Information Science and Technology, 2007. 58(7): p. 1019-1031.

[14] Goyal, S., M.J. van der Leij, and J.L. Moraga-Gonzalez, Economics: An emerging small world. Journal of Political Economy, 2006. 114(2): p. 403-412.

[15] Radicchi, F., et al., Diffusion of scientific credits and the ranking of scientists. Physical Review E, 2009. 80(5).

[16] Yoshikane, F., et al., An analysis of the connection between researchers' productivity and their co-authors' past attributions, including the importance in collaboration networks. Scientometrics, 2009. 79(2): p. 435-449.

[17] Wagner, C.S. and L. Leydesdorff, Network structure, self-organization, and the growth of international collaboration in science. Research Policy, 2005. 34(10): p. 1608-1618.

[18] He, B., Y. Ding, and C. Ni, Mining Enriched Contextual Information of Scientific Collaboration: A Meso Perspective. Journal of the American Society for Information Science and Technology, 2011. 62(5): p. 831-845.

[19] Milojević, S., Accuracy of simple, initials-based methods for author name disambiguation. Journal of Informetrics, 2013. 7(4): p. 767-773.

[20] Fegley, B.D. and V.I. Torvik, Has Large-Scale Named-Entity Network Analysis Been Resting on a Flawed Assumption? Plos One, 2013. 8(7).

[21] Kim, J., H. Kim, and J. Diesner, The Impact of Name Ambiguity on Properties of Coauthorship Networks. Journal of Information Science Theory and Practice, 2014. 2(2): p. 6-15.

[22] Cavero, J.M., B. Vela, and P. Caceres, Computer science research: more production, less productivity. Scientometrics, 2014. 98(3): p. 2103-2111.

[23] Shi, Q., et al., Diversity of social ties in scientific collaboration networks. Physica a-Statistical Mechanics and Its Applications, 2011. 390(23-24): p. 4627-4635.

[24] Ley, M., The DBLP Computer Science Bibliography: Evolution, Research Issues, Perspectives, in String Processing and Information Retrieval, A.F. Laender and A. Oliveira, Editors. 2002, Springer Berlin Heidelberg. p. 1-10.

[25] Ley, M., DBLP: some lessons learned. Proc. VLDB Endow., 2009. 2(2): p. 1493-1500.

[26] Deville, P., et al., Career on the Move: Geography, Stratification, and Scientific Impact. Scientific Reports, 2014. 4.

[27] Eom, Y.-H. and H.-H. Jo, Generalized friendship paradox in complex networks: The case of scientific collaboration. Scientific reports, 2014. 4: p. 4603.

[28] Torvik, V.I. and N.R. Smalheiser, Author Name Disambiguation in MEDLINE. Acm Transactions on Knowledge Discovery from Data, 2009. 3(3).

[29] Kim, S. and S. Cho, Characteristics of Korean Personal Names. Journal of the American Society for Information Science and Technology, 2013. 64(1): p. 86-95.