Journal of Informetrics 9 (2015) 226–236

Contents lists available at ScienceDirect

Journal of Informetrics

journal homepage: www.elsevier.com/locate/joi

The effect of data pre-processing on understanding the evolution of collaboration networks

Jinseok Kim*, Jana Diesner

Graduate School of Library and Information Science, University of Illinois at Urbana-Champaign, 501 East Daniel Street, Champaign, IL 61820, USA

ARTICLE INFO

Article history:
Received 21 September 2014
Received in revised form
28 December 2014
Accepted 5 January 2015
Available online 17 January 2015

Keywords:
Collaboration network
Network evolution
Name ambiguity
Disambiguation

ABSTRACT

This paper shows empirically how the choice of certain data pre-processing methods for disambiguating author names affects our understanding of the structure and evolution of co-publication networks. Thirty years of publication records from 125 Information Systems journals were obtained from DBLP. Author names in the data were pre-processed via algorithmic disambiguation. We applied the commonly used all-initials and first-initial based disambiguation methods to the data, generated over-time networks with a yearly resolution, and calculated standard network metrics on these graphs. Our results show that initial-based methods underestimate the number of unique authors, average distance, and clustering coefficient, while overestimating the number of edges, average degree, and ratios of the largest components. These self-reinforcing growth and shrinkage mechanisms amplify over time. This can lead to false findings about fundamental network characteristics such as topology and reasoning about underlying social processes. It can also cause erroneous predictions of trends in future network evolution and suggest unjustified policies, interventions and funding decisions. The findings from this study suggest that scholars need to be more attentive to data pre-processing when analyzing or reusing bibliometric data.

1. Introduction and background

The growth of scholarly collaboration networks has recently attracted the attention of different scholarly communities. For example, Barabási et al. (2002) modeled how authors choose coauthors based on the prior number of collaborators over a 7-year time window and showed that this mechanism can lead to a degree distribution with a slope following a power law. Based on bibliometric data spanning from 1960 to 2008, Franceschet (2011) presented how collaboration networks in computer science have "lost peculiar core-periphery structure over time" (p. 2009). Such studies on evolving networks have been carried out in various academic subfields: e.g., mathematics (Barabási et al., 2002; Grossman, 2002), neuroscience (Barabási et al., 2002), and physics (Lee, Goh, Kahng, & Kim, 2010). Some scholars have tracked growing coauthorship networks at a national level: e.g., for Japan (Yoshikane & Kageura, 2004), Slovenia (Perc, 2010) and Turkey (Çavuşoğlu & Türker, 2013).

* Corresponding author. Tel.: +1 217 751 2943; fax: +1 217 244 3302.
E-mail address: jkim362@illinois.edu (J. Kim).

http://dx.doi.org/10.1016/j.joi.2015.01.002
1751-1577/

Studies of collaboration network evolution mostly construct network data based on bibliometric records. Two people (i.e., network nodes) are connected by a coauthoring relationship (i.e., edges) if they appear as coauthors in the byline of a paper. Network construction implies the following challenge: an author's identity is usually represented by a name string in the raw data. This representation can be a source of name ambiguity. For example, two given name instances, 'Black, Samuel' and 'Black, Samuel' may sometimes refer to the same person, but other times to different people who happen to have the same name (i.e., homonym). Another situation that can lead to errors is when 'Black, Samuel' is the same person as 'Black, S.' if the same author used a first name initial in one paper but a full given name in another publication (i.e., synonym).

To address the problem of name ambiguity in bibliometric data, some scholars have used datasets where name ambiguity has already been resolved by data providers: e.g., the Digital Bibliography & Library Project (DBLP) (Franceschet, 2011) or Mathematical Reviews (Grossman, 2002). Others have devised their own methods for disambiguating raw datasets: e.g., the Physical Review Series published by the American Physical Society (Deville et al., 2014; Martin, Ball, Karrer, & Newman, 2013) and Italian scholars' publications from ISI Web of Science (Abramo, D'Angelo, & Murgia, 2013).

Such disambiguated data are, however, limited in coverage of fields and availability. Thus, the simple heuristic that an author can be represented by a full surname and given-name initials has been widely used (Milojević, 2013; Strotmann & Zhao, 2012). According to this heuristic, author names are assumed to refer to the same person if they share the initial of the first name (i.e., first-initial method hereafter) or all the initials of first and middle names (i.e., all-initials method hereafter). Scholars have well acknowledged that this data pre-processing decision may entail errors of misidentification (e.g., Barabási et al., 2002; Newman, 2001). The names of different authors may sometimes be merged into one author identity. For example, 'Black, S. (actually S. for Samuel)' can be regarded as one with 'Black, S. (actually S. for Susan)' as both share the first letter in given names. Another type of error occurs where names of the same author are split into different identities such as 'Black, S.' and 'Black, S. L.' when an author inconsistently uses her/his middle name initial.

Despite this possibility of misidentification, the use of initial-based disambiguation in bibliometric data has been supported for several practical reasons. First, a majority of names in many bibliometric datasets come in the format of a full surname followed by given name initial(s) (e.g., ISI Web of Science[1] or SCOPUS). Additionally, even sophisticated disambiguation algorithms do not guarantee perfect disambiguation (Wagner & Leydesdorff, 2005). Most importantly, misidentification errors are not necessarily assumed to be critical to research outcomes. In other words, findings from networks disambiguated by initials are believed to approximate the network properties of ground-truth data quite accurately (e.g., Barabási et al., 2002; Milojević, 2013; Newman, 2001).

The proposition of the accuracy of initial-based disambiguation has been recently tested by several scholars. For example, Fegley and Torvik (2013) showed that 3.2 million unique authors in algorithmically disambiguated MEDLINE data can be reduced to 1.6 million by the first-initial disambiguation, and several network properties such as degree distribution and clustering coefficient can thereby be distorted. These findings were confirmed based on DBLP data by Kim, Kim, and Diesner (2014). Such a distortive effect of name ambiguity was, however, shown to decrease to a negligible extent when only last-positioned author names in bylines are considered (Strotmann & Zhao, 2012).

As such, this paper is first motivated by the fact that the accuracy of initial-based disambiguation and its impact on network properties are still disputable. Interestingly, for example, two different methods have been applied to raw data from the same source, e.g., for the Physical Review Series data from the American Physical Society, researchers have employed algorithmic disambiguation (Deville et al., 2014; Martin et al., 2013) vs. all-initials method (Eom & Jo, 2014; Radicchi, Fortunato, Markines, & Vespignani, 2009). Moreover, the impact of the data pre-processing method on research findings has rarely been discussed in the context of network evolution. Aforementioned studies that test the performance of initial-based disambiguation have mainly focused on a static view of collaboration network structure.

In this sense, we believe this paper expands the works by Fegley and Torvik (2013) and Kim et al. (2014) by investigating a temporal aspect of network formation. In addition, our approach is different in that two previous exemplar studies were based on the exceptionally large-scale data: 2 million papers in Fegley and Torvik's and 1 million papers in Kim et al.'s. It is possible that the impact of name ambiguity can become negligible if a target dataset is smaller than those of two preceding studies. Thus, we selected a dataset of 113,000 publication records that are similar to or smaller than those used in previous evolutionary coauthorship network studies. Moreover, we also attempt to address how the choice of disambiguation methods can lead to different network topologies, which has not been directly discussed in previous studies.

Therefore, the purpose of this study is to contribute to the discussion of the effect of name ambiguity on research findings by demonstrating how the selection of data pre-processing methods can affect the representation of evolving network properties. Our paper does not attempt to refute or raise questions about previous studies, nor do we take sides on any specific disambiguation method. Instead, this study is expected to serve as an example to motivate readers to pay more attention to the importance of data pre-processing in bibliometric research. In the following section, the choice of data and measurements are explained.

[1] It should be noted that ISI Web of Science provides full given names, when available, for many of publication records but usually for a recent period, e.g., 2006 and afterwards.

2. Methodology

2.1. Data

The data were obtained from the Digital Bibliography & Library Project (DBLP hereafter) database. DBLP is well known for its name disambiguation: author names in DBLP are recorded in full name formats, if available, and disambiguated by algorithms utilizing mainly coauthor information and also by manual inspection of suspicious name pairs during the data pre-processing stage (Ley, 2009). Publication records in DBLP can be freely downloaded in XML format. We also directly retrieved the list of synonym pairs of authors from the DBLP webpage. Following the example of prior studies (e.g., Cavero, Vela, & Caceres, 2014), we considered this list during the data parsing.

From the parsed DBLP data, we retrieved publication records from 125 journals indexed by the 2013 Journal Citation Report for the 'Information Systems' field. We chose this subfield of computer science as it has been included in or a target field of many collaboration network studies (e.g., Cavero et al., 2014; Fiala, 2012; Franceschet, 2011). The selected data consists of a total of 113,038 journal publications spanning from 1984 to 2013. As we are interested in collaboration networks, only papers with two or more authors were further considered.

As the aim of this paper is to compare the effect of name disambiguation methods on properties of evolving network, the data were disambiguated by all-initials and first-initial methods as described in Milojević (2013) to produce three versions of the dataset for network generation. The first network was constructed from the DBLP raw data (previously disambiguated by DBLP). The second one was generated from the same data but disambiguated by the all-initials method, while the third one was disambiguated by the first-initial method. As the original DBLP data recorded an author name in the format of a given name followed by a surname, we converted each name instance into the surname followed by a given name format, using the steps described in Kim et al. (2014).

We used a baseline to compare the results from our experiments against. More specifically, we obtained a ground-truth dataset of 474 unique authors who have highly common i.e. ambiguous names such as Smith, Wang, or Kumar in 3,921 DBLP publication records. The ground-truth dataset was generated by Shin, Kim, Choi, and Kim (2014). We calculated the accuracy of DBLP's algorithmic disambiguation, all-initials method, and first-initial method against the ground-truth in terms of *K*-metric and pairwise F1. These metrics have been widely used by name disambiguation scholars to test the performance of disambiguation methods (Ferreira, Gonçalves, & Laender, 2012).

K-metric: This is the geometric mean of average cluster purity (ACP) and average author purity (AAP).

$$K = \sqrt{\mathrm{ACP} \times \mathrm{AAP}}$$

$$\mathrm{ACP} = \frac{1}{N}\sum_{i=1}^{q}\sum_{j=1}^{R}\frac{n_{ij}^2}{n_i}$$

$$\mathrm{AAP} = \frac{1}{N}\sum_{j=1}^{R}\sum_{i=1}^{q}\frac{n_{ij}^2}{n_j}$$

In the equations, N is the sum of name instances; R is the number of ground-truth clusters; q is the number of clusters generated by algorithmic disambiguation of DBLP or initial-based disambiguation; n_{ij} is the number of elements of cluster i in q belonging to the cluster j in R; n_i and n_j represent the number of elements in the cluster i and j.

If all clusters contain only the correct name instances belonging to the same identities, then the ACP value will be 1. The ACP value decreases if clusters include merged identities (high merging). Meanwhile, if each cluster has a small number of name instances that should belong to this cluster but is not included in it (low splitting), the AAP value gets closer to 1.

Pairwise F1: Pairwise precision (pP) is calculated as $p\mathrm{P} = A/(A+C)$, while pairwise recall (pR) is calculated as $p\mathrm{R} = A/(A+B)$. Here, A is the number of pairwise name instances in clusters generated by algorithmic or initial-based disambiguation methods that are correctly assigned to the same authors (=true positives), while C is the number of pairwise name instances in the clusters but do not belong to the same authors (=false positives). B is the number of pairwise name instances that are associated with the same authors but are not included in the disambiguated clusters (=false negatives). From these two metrics, the pF1 is defined as follows:

$$p\mathrm{F1} = \frac{(\beta^2+1)\times p\mathrm{P}\times p\mathrm{R}}{\beta^2\times p\mathrm{P} + p\mathrm{R}}$$

Here, β is the weight of recall relative to precision. We use $\beta = 1$ as in F1, which weighs two metrics equally.

In Table 1, the algorithmic disambiguation by DBLP is shown to outperform initial-based methods. Although the DBLP disambiguation method could not perfectly disambiguate names, its performance is similar to or slightly higher than other advanced algorithms in previous studies (e.g., Cota, Ferreira, Nascimento, Goncalves, & Laender, 2010; Pereira et al., 2009).

Table 1
Accuracy of disambiguation methods and number of unique authors identified by each method.

Method	*K*-metric	Pairwise F1	Unique authors
Algorithmic	0.952	0.96	448
All-initials	0.643	0.33	149
First-initial	0.284	0.09	18

Thus, we use the DBLP data as a baseline against which the performance of initial-based disambiguation is compared to in terms of network properties.

2.2. Measurements

We consider the following set of network metrics as defined in Newman (2001) and commonly used in co-publishing network studies. Each metric was calculated by *Pajek* (version 4.01).

Number of unique authors: The total number of unique authors identified by each method, corresponding to unique nodes in coauthorship networks. In our DBLP data, a disambiguated author identity is represented by a unique name string that usually consists of a full surname and given name format. If two or more authors have the same name string, a numeric code of four digits is assigned to differentiate each author (e.g., Jun Wang, Jun Wang 0001, Jun Wang 0002, etc.). When disambiguated by initial methods, an author identity is represented by a full surname followed by the first initial or all initials of a given name. For example, 'Black, Samuel Loy' is represented as 'Black, S.' by first-initial disambiguation, while as 'Black, S. L.' by the all-initials method.

Number of edges: In collaboration networks, an edge represents the coauthoring relationship between two authors. Here, we do not consider frequency or strength of collaboration (Barabási et al., 2002; Newman, 2001). In other words, the existence or absence of relations is the only criterion that matters for this metric.

Degree: In collaboration networks, this means the number of unique coauthors an author has ever worked with.

Components: A component of a network is a subset of nodes in which any pair of nodes can reach one another. As most collaboration networks have many disconnected components, scholars are usually interested in the size of the largest component and its ratio over the whole network (Liu, Bollen, Nelson, & Van de Sompel, 2005).

Distance: The length of the shortest paths or geodesics between any two authors in the same component. We calculated the mean distance of all reachable pairs of authors in collaboration networks (Brandes, 2001).

Clustering coefficient: The clustering coefficient measures the density (i.e. number of existing edges over number of possible edges) of each node's ego networks and averages these values for the entire network. This is often approximated as the ratio of the number of triangles over the number of triples connected with two edges (Newman, 2001). The latter operationalization translates into the probability that a pair of scholars who both have collaborated with a common third scholar tend to work together (Franceschet, 2011).

3. Analysis

3.1. Number of unique authors

Fig. 1 shows the temporal change in the numbers of unique authors depending on each of the three data pre-processing methods: algorithmically disambiguated (blue circles), all-initials (black crosses), and first-initial (red triangles). The number

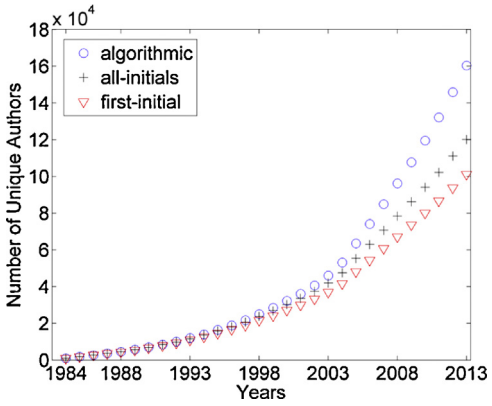

Fig. 1. Trend of number of unique author (cumulative up to the indicated year). (For interpretation of the references to color in the text, the reader is referred to the web version of this article.)

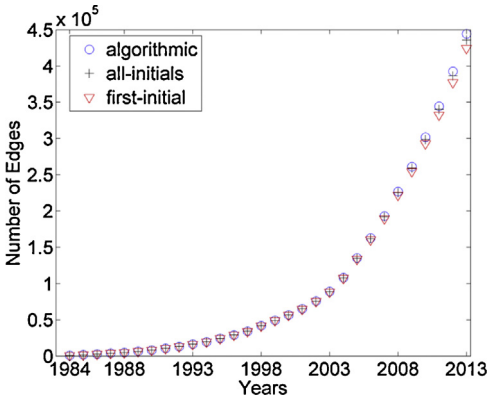

Fig. 2. Trend of number of edges (cumulative up to the indicated year).

of unique authors up to the target year is plotted by a yearly resolution for each method. In all three plots, the number of unique authors shows an exponential growth over the years: disambiguated ($y = 1986.5e^{0.156x}$, $R^2 = 0.97$), all-initials ($y = 2104.9e^{0.147x}$, $R^2 = 0.96$), and first-initial ($y = 2097.8e^{0.141x}$, $R^2 = 0.96$). This indicates that all three data pre-processing methods suggest the same type of growth rate.

One noticeable observation, however, is that the number of unique authors is consistently underestimated by initial-based methods for all years, although in Fig. 1 the three plots seem to overlap in early periods. This means that, through initial-based disambiguation, merging of author identities happens more often than splitting (merging reduces the number of unique identities while splitting increases it). Moreover, this finding can counter the assumption that the all-initials method can provide an upper limit of the 'true' number of unique authors, while the first-initial method provides the lower limit (Newman, 2001). Based on this assumption, scholars have argued that the statistical property of a perfectly disambiguated network can be estimated to lie between network properties disambiguated by the all-initials and first-initial methods (e.g., Barabási et al., 2002; Newman, 2001).

Another observation is that the gaps between plot lines have continued to increase over time. For example, in 1984, DBLP identified 962 authors, while the all-initials method found 961 and first-initial found 957. In 2013, these numbers are 160,349 by algorithmic disambiguation of DBLP, 120,052 authors by all-initials (−25.13% in comparison to algorithmic disambiguation) and 101,003 by first-initial (−37.01% in comparison to algorithmic disambiguation). The differences in the numbers of unique author depending on the chosen data processing method show approximately an exponential growth: disambiguated vs. all-initials ($y = 8.45e^{0.308x}$, $R^2 = 0.95$), and disambiguated vs. first-initial ($y = 40.65e^{0.267x}$, $R^2 = 0.93$). This implies that the prediction of the size of a scientific community in the target IS journals in the future can be very different according to the disambiguation method that is used.

3.2. Number of edges

Similar to the trend for unique author numbers, the numbers of edges show an exponential growth over time for all three disambiguation techniques. Unlike the unique author number trend, however, the plots of edges show no big gaps between them. As shown in Fig. 2, the numbers of edges overlap in most years and their exponential trend lines have similar coefficients for modeling growth: disambiguated ($y = 1795.4e^{0.1949x}$, $R^2 = 0.98$), all-initials ($y = 1798.4e^{0.1945x}$, $R^2 = 0.98$), first-initial ($y = 1806.6e^{0.1938x}$, $R^2 = 0.98$). The biggest gap is found in 2013 between the number of edges (444,411) in the algorithmically disambiguated network and the number of edges (424,287, −4.53%) based on the first-initial method.

This finding indicates that merged author nodes usually have distinct collaborators. If two merged authors have coauthors that are also merged because of their shared first or middle name initials, then the edges between each merged author and her/his coauthor would also be consolidated into one edge. If this merging of edges happens frequently, then the total number of edges in the network would decrease to a noticeable extent. As this decrease happens at a minimal level, we can infer that, in our dataset, it is uncommon that two or more authors in a byline have ambiguous names that may lead to merging with names in other bylines.

3.3. Degree

In Fig. 3, the average degrees of authors identified by each method are plotted in the same way as described for Fig. 1 above. Two plots show an exponential growth with very low slopes: all-initials ($y = 1.71e^{0.048x}$, $R^2 = 1.00$) and first-initial ($y = 1.72e^{0.053x}$, $R^2 = 1.00$), while the degree plot of algorithmically disambiguated network grows almost linearly. Contrary to the number of unique authors, the average degrees by initial-based methods are found to be overestimated. For example, as of 2013, the average degree of authors in the algorithmically disambiguated dataset is 5.54, which increases to 7.26 (30.99%) with the all-initials method and to 8.40 (51.57%) with the first-initial method. This phenomenon can be explained mainly

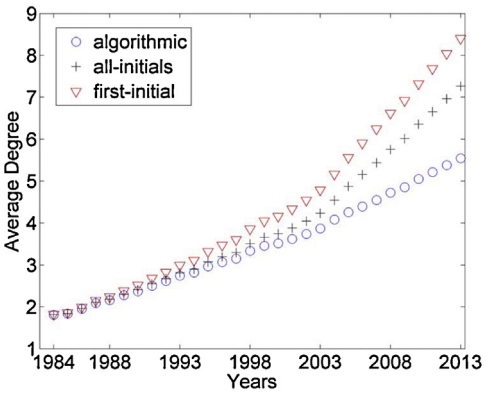

Fig. 3. Trend of average degree (cumulative up to the indicated year).

by the merging of author identities. When two distinct author identities are merged into one, their coauthoring partners are also attached to the merged identity; increasing the number of collaborators (i.e., degree). While merged authors become connected to more collaborators, the number of unique authors decreases due to the merging. These two effects erroneously inflate the average degree.

This figure also shows that gaps between plots have increased over time and we can conjecture that these gaps will continue to increase in the future if the current trend continues. The gap between initial methods, namely all-initials vs. first-initial, is smaller than those between algorithmically disambiguation vs. all-initials or vs. first-initial method. However, all those gaps show approximately an exponential growth: disambiguated vs. all-initials ($y = 0.006e^{0.206x}$, $R^2 = 0.96$) and disambiguated vs. first-initial ($y = 0.033e^{0.165x}$, $R^2 = 0.93$).

With regard to degree, scholars have often tested degree distribution to determine whether the topology of the target network can be characterized as a power-law network (Barabási et al., 2002; Milojević, 2010; Newman, 2001). When the power-law slope was found to characterize the network, plausible mechanisms leading to such a distribution were also proposed and tested. One of those mechanisms is preferential attachment, i.e. authors' natural inclination to choose collaborators who already have many collaborators (Barabási et al., 2002). Such a tendency was believed to accumulate over time in an evolving collaboration network; finally generating the power law distribution of degree in networks.

Such an evolution-generates-topology relationship can be tested by comparing the time series representations of degree distributions, as shown in Fig. 4. In that figure, the degree distributions of authors identified by three data-preprocessing methods are shown in cumulative log–log plots as measured over a 5 year windows from 1984 to 2013. Each of the six subfigures contains the cumulative distribution plots up to the target year.

In these figures, the x-axis represents the value of degree (x), while the y-axis scales the proportion of authors who have the target value ($X = x$) and above ($X > x$). First, the results suggest that distributions obtained with first-initial method are positioned above those of all-initials method, which in turn are positioned above those of algorithmic disambiguation. This means that, for a given x degree, the number of authors who have a degree of that value or higher tends to be inflated by initial-based methods. This is in line with the prior findings in this section. As authors are merged by initial-based methods, their collaborators are attached to the merged identities. This merging effect pushes the plots right and upward when compared to those created using algorithmic disambiguation.

The plot gaps resulting from these merging effect phenomenon get wider over time as the merged identities increase every year (see Fig. 1). Such a temporal increase of compromised author identities and its subsequent effect on degrees of authors generate different distribution plots by each method. As the collaboration networks evolve over time, some degree distributions may be roughly fitted to a power law distribution with cutoffs. For example, the degree distributions in 2013 by initial-based methods show slopes roughly approximating a power law although the fits are not statistically significant: slope = −2.82 ($x \geq 27$, $p = 0.018$) by all-initials method and slope = −2.65 ($x \geq 27$, $p = 0.043$) by first-initial method. Here, we fitted our data to power law distribution as described in Clauset, Shalizi, and Newman (2009).[2] This implies that, depending on data pre-processing methods, the same network can be characterized by different topologies and generation mechanisms – independently of any actual network effects,[3] but purely induced by pre-processing steps.

To illustrate this propagation of errors over time further, we generated three simulated scale-free networks based on the preferential attachment mechanism (Barabási et al., 2002). The simulated networks have the same or similar number

[2] According to their method, p values above 0.1 are interpreted to indicate a statistical significance. For a detailed explanation on the meaning of p in power-law fitting, we refer the reader to Clauset et al. (2009).

[3] This statement assumes that the distribution of a network disambiguated by algorithm serves as a baseline. If the distribution of networks pre-processed by initial-based methods was set as a baseline, the distribution plot of algorithmic disambiguation would be possibly a case of a false negative finding, i.e., no power-law distribution.

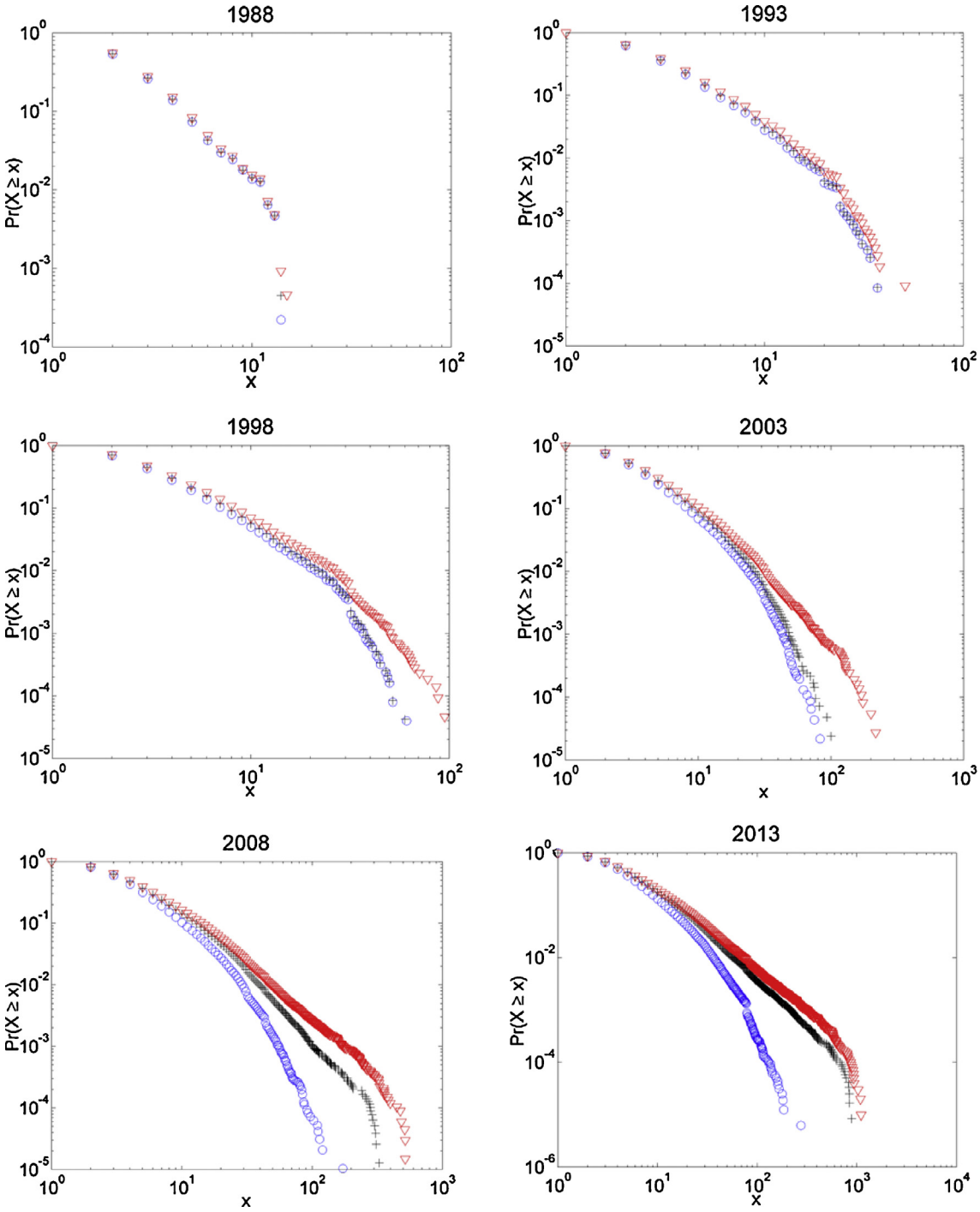

Fig. 4. Cumulative log–log plot of degree distributions per period (blue circle = algorithmic, black cross = all-initials and red triangle = first-initial). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

of unique coauthors, edges, and average degrees as the networks resulting from algorithmic, all-initials, and first-initial disambiguation (see Table 2).[4] Each cumulative log–log plot of their degree distributions (black solid line) is shown in Fig. 5, along with those by algorithmic (blue circles), all-initials (black crosses), and first-initial (red triangles) methods.

[4] *Pajek* (de Nooy, Mrvar, & Batagelj, 2011), a network analysis package, was used to generate scale-free networks with same parameters applied: number of vertices in initial networks (10), initial probability of edges (0.20), and alpha (0.25).

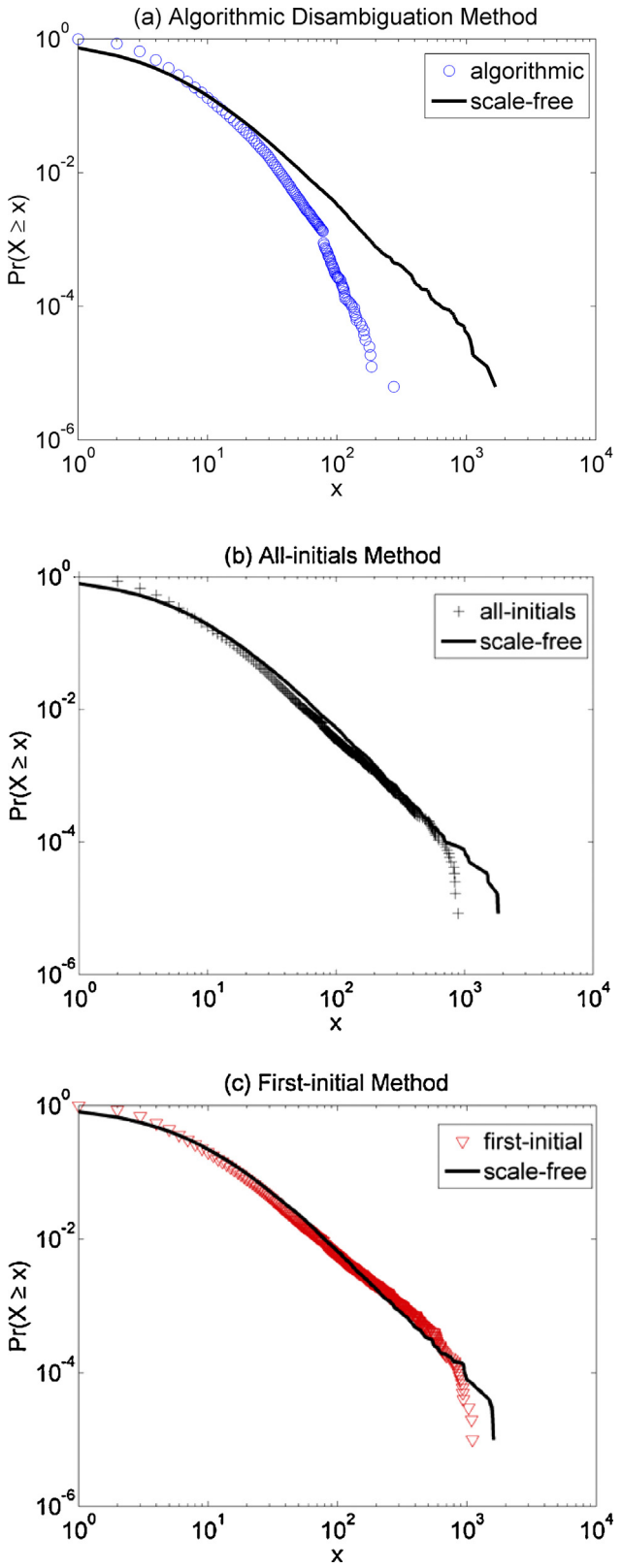

Fig. 5. Cumulative log–log plot of degree distribution of scale-free networks corresponding to networks by algorithmic, all-initials, and first-initial disambiguation. (For interpretation of the references to color in the text, the reader is referred to the web version of this article.)

Table 2
Summary of scale-free networks per disambiguation method (values in parentheses from corresponding actual networks).

Disambiguation method	No. of unique author	No. of edge	Avg. degree
Algorithmic	160,349 (160,349)	443,500 (444,411)	5.53 (5.54)
All-initials	120,052 (120,052)	433,564 (435,830)	7.22 (7.26)
First-initial	101,003 (101,003)	421,815 (424,287)	8.35 (8.40)

Interestingly, the distribution plots from scale-free networks seem to come close to the slopes obtained from the networks with all-initials and first-initial methods across many x values.

3.4. Components

Fig. 6 tracks the ratios of the size of the largest component over time. Overall, the ratio plots of the largest components of networks disambiguated by algorithm are located below the plots of networks processed by initial-based methods. This means that networks with initial-based name processing tend to inflate the ratios of the largest components in comparison to those of algorithmically disambiguated networks. This can be explained partly by the merging effect. When author identities are merged into other ones in a network, they also attach their local networks to the merged identities, which leads to an increase in the size of the largest component.

The gap of ratios between algorithmically disambiguated data and initial-based processed data increased for some time and then moderately decreased. For example, in 1997, the ratio of the largest component in the network from first-initial method was 67.29%, while that by algorithmic disambiguation was 40.13%. The gap (27.16%p) began to decrease afterwards to reach a point in 2013 where the former was 87.87% and the latter was 75.98% (the gap was 11.89%p). The observed fluctuation of gap size can be explained by structural characteristics of incorrectly merged authors. If many of the authors who are merged by initial-based methods happen to be in the same component, the increase of the component size would not be noticeable compared to the situation when they are in separate components or isolated from components before merging. Before 1999, therefore, many of merged authors seemed to attach their isolated local networks to the largest component; increasing its ratio, while after then such an attachment by merging seemed to weaken.

3.5. Distance and clustering coefficient

As networks evolve over time, the average distances between authors increased until around 1990 and then decreased regardless of data pre-processing methods (see Fig. 7). Such a trend was confirmed in prior network evolution studies (e.g., Franceschet, 2011; Perc, 2010). After 1994, the mean distance of the network disambiguated by algorithms keeps showing higher values than those of networks pre-processed by initial-based methods. This can be explained by considering that merged authors in networks where authors are identified by given-name initials act as bridges connecting authors who were unreachable, or by providing shorter paths for authors who were reachable with longer paths. This process is regarded to reduce the average distance (7.32) of the algorithmically disambiguated network to 5.54 (−24.38%) by the all-initials method and to 5.14 (−29.77%) by the first-initial method as of 2013.

Clustering coefficients of networks, as displayed in Fig. 8, show an overall decreasing trend. The gaps between networks pre-processed by these three methods become wider over time. This can be explained as follows: the clustering coefficient in our study is calculated as the proportion of triangles (i.e. three nodes being all connected to each other) over triples connected with two edges (i.e., possible triangles). When authors' identities are merged, the number of triples (denominator) increases. During this merging process, however, the number of triangles (numerator) may not increase at a corresponding rate. For example, when two authors are connected via a merged author, a triple forms between them. If they have not actually

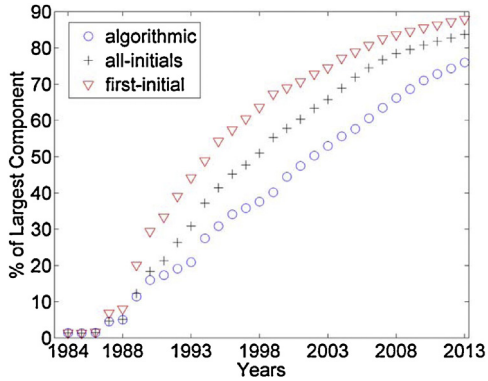

Fig. 6. Trend of ratios of the largest component (cumulative up to the indicated year).

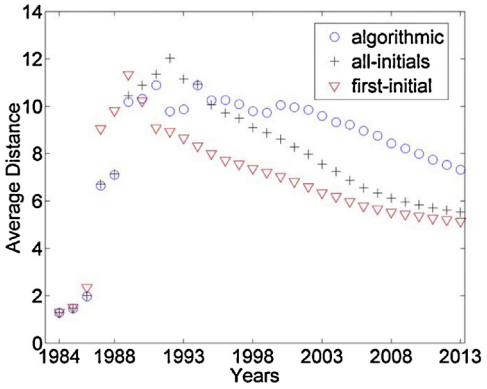

Fig. 7. Trend of average distance (cumulative up to the indicated year).

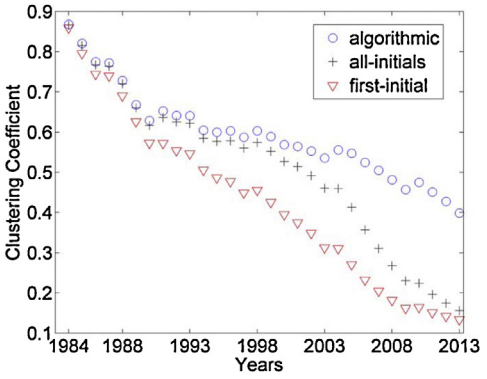

Fig. 8. Trend of clustering coefficient (cumulative up to the indicated year).

collaborated, a triangle fails to form. If this happens often when authors are merged, the clustering coefficient of the network begins to decrease. As a result, despite the common trend that the clustering coefficient decreases, networks pre-processed by each method can provide different findings in terms of magnitude of change. For example, in 2013, the clustering coefficient of the algorithmically disambiguated network is 0.398, which is reduced to 0.156 (−60.89%) by all-initials method and 0.134 (−66.42%) by first-initial method.

4. Conclusion and discussion

This paper illustrates how certain choices for pre-processing network data can affect our knowledge about the properties and evolution of collaboration networks and reasoning about underlying social processes. To study the magnitude of these effects, a bibliometric dataset was obtained from DBLP where author names had been algorithmically disambiguated in a highly accurate fashion. We applied all-initials and first-initial based disambiguation – two commonly used pre-processing techniques in bibliometrics – to these data. We generated three over-time networks that each represent one of these three choices (disambiguation based on algorithms, first initials, and all-initials, respectively), calculated commonly used network metrics, and compared the temporal changes in these metrics within and across the three graphs.

Overall, the evolution of networks pre-processed by these three methods showed similar trends, namely: an exponential growth in the number of authors, number of edges and average degree. However, the fine-grained analysis revealed that certain network properties and value ranges differ substantially depending on pre-processing choices: in comparison to the algorithmically disambiguated data, which is as close to ground-truth as we can get it, the graphs pre-processed with initial-based methods underestimated the number of unique authors, average distance, and clustering coefficient, while overestimating the number of edges (at a small level), average degree, and ratios of the largest components. These self-reinforcing growth and shrinkage rates became intensified over time.

The findings from this study suggest that scholars should be attentive to data pre-processing choices when analyzing, curating or reusing bibliometric data. In extreme cases, as shown with regard to degree distribution, a certain data pre-processing method may lead to false conclusions about the network topology and its generation mechanism. Moreover, the erroneous changes in network properties over time can lead to false predictions of network evolution into the future; potentially affecting policy and funding decisions.

This study entails several limitations. First, the findings should be tested and confirmed against other subfields of computer and information science, and other disciplines. Second, the impact of data pre-processing on growing network properties considering varying data sizes should be investigated. It would be helpful to identify the level of name ambiguity in a network data, which can reduce the differences due to data pre-processing methods to a negligible extent. Furthermore, the process of how merged and/or split author identities affect the network structures needs to be explored. These considerations are left for future research.

Acknowledgments

This work was supported by KISTI (Korea Institute of Science and Technology Information), grant P14033. We would like to thank Jose Maria Cavero, Belen Vela, and Paloma Caceres for helping us to match IS journal names in DBLP and ISI Web of Science, and Natalie Lambert (University of Illinois at Urbana-Champaign) for editing the manuscript.

References

Abramo, G., D'Angelo, C. A., & Murgia, G. (2013). The collaboration behaviors of scientists in Italy: A field level analysis. *Journal of Informetrics*, *7*(2), 442–454. http://dx.doi.org/10.1016/j.joi.2013.01.009

Barabási, 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. http://dx.doi.org/10.1016/s0378-4371(02)00736-7

Brandes, U. (2001). A faster algorithm for betweenness centrality. *Journal of Mathematical Sociology*, *25*(2), 163–177.

Cavero, J. M., Vela, B., & Caceres, P. (2014). Computer science research: More production, less productivity. *Scientometrics*, *98*(3), 2103–2111. http://dx.doi.org/10.1007/s11192-013-1178-2

Çavuşoğlu, A., & Türker, İ. (2013). Scientific collaboration network of Turkey. *Chaos, Solitons & Fractals*, *57*, 9–18.

Clauset, A., Shalizi, C. R., & Newman, M. E. J. (2009). Power-law distributions in empirical data. *SIAM Review*, *51*(4), 661–703.

Cota, R. G., Ferreira, A. A., Nascimento, C., Goncalves, M. A., & Laender, A. H. F. (2010). An unsupervised heuristic-based hierarchical method for name disambiguation in bibliographic citations. *Journal of the American Society for Information Science and Technology*, *61*(9), 1853–1870. http://dx.doi.org/10.1002/asi.21363

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

Deville, P., Wang, D., Sinatra, R., Song, C., Blondel, V. D., & Barabási, A. L. (2014). Career on the move: Geography, stratification, and scientific impact. *Scientific Reports*, *4* http://dx.doi.org/10.1038/srep04770

Eom, Y.-H., & Jo, H.-H. (2014). Generalized friendship paradox in complex networks: The case of scientific collaboration. *Scientific Reports*, *4*, 4603. http://dx.doi.org/10.1038/srep04603

Fegley, B. D., & Torvik, V. I. (2013). Has large-scale named-entity network analysis been resting on a flawed assumption? *PLoS ONE*, *8*(7) http://dx.doi.org/10.1371/journal.pone.0070299

Ferreira, A. A., Gonçalves, M. A., & Laender, A. H. F. (2012). A brief survey of automatic methods for author name disambiguation. *Sigmod Record*, *41*(2), 15–26.

Fiala, D. (2012). Time-aware PageRank for bibliographic networks. *Journal of Informetrics*, *6*(3), 370–388. http://dx.doi.org/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. http://dx.doi.org/10.1002/asi.21614

Grossman, J. W. (2002). Patterns of collaboration in mathematical research. *SIAM News*, *35*(9), 8–9.

Kim, J., Kim, H., & Diesner, J. (2014). The impact of name ambiguity on properties of coauthorship networks. *Journal of Information Science Theory and Practice*, *2*(2), 6–15. http://dx.doi.org/10.1633/JISTaP.2014.2.2.1

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

Ley, M. (2009). DBLP: Some lessons learned. *Proceedings of the VLDB Endowment*, *2*(2), 1493–1500.

Liu, X. M., Bollen, J., Nelson, M. L., & Van de Sompel, H. (2005). Co-authorship networks in the digital library research community. *Information Processing & Management*, *41*(6), 1462–1480. http://dx.doi.org/10.1016/j.ipm.2005.03.012

Martin, T., Ball, B., Karrer, B., & Newman, M. E. J. (2013). Coauthorship and citation patterns in the Physical Review. *Physical Review E*, *88*(1) http://dx.doi.org/10.1103/PhysRevE.88.012814

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. http://dx.doi.org/10.1002/asi.21331

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

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. http://dx.doi.org/10.1073/pnas.021544898

Perc, M. (2010). Growth and structure of Slovenia's scientific collaboration network. *Journal of Informetrics*, *4*(4), 475–482.

Pereira, D. A., Ribeiro-Neto, B., Ziviani, N., Laender, A. H. F., Gonçalves, M. A., & Ferreira, A. A. (2009). Using web information for author name disambiguation. In *Paper presented at the Proceedings of the 9th ACM/IEEE-CS joint conference on Digital libraries* Austin, TX, USA.

Radicchi, F., Fortunato, S., Markines, B., & Vespignani, A. (2009). Diffusion of scientific credits and the ranking of scientists. *Physical Review E*, *80*(5) http://dx.doi.org/10.1103/PhysRevE.80.056103

Shin, D., Kim, T., Choi, J., & Kim, J. (2014). Author name disambiguation using a graph model with node splitting and merging based on bibliographic information. *Scientometrics*, *100*(1), 15–50. http://dx.doi.org/10.1007/s11192-014-1289-4

Strotmann, A., & Zhao, D. (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. http://dx.doi.org/10.1002/Asi.22695

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

Yoshikane, F., & Kageura, K. (2004). Comparative analysis of coauthorship networks of different domains: The growth and change of networks. *Scientometrics*, *60*(3), 435–446. http://dx.doi.org/10.1023/b:scie.0000034385.05897.46