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.