THE IMPACT OF AUTHOR NAME DISAMBIGUATION ON KNOWLEDGE DISCOVERY FROM LARGE-SCALE SCHOLARLY DATA

BY

JINSEOK KIM

DISSERTATION

Submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy in Library and Information Science
in the Graduate College of the
University of Illinois at Urbana-Champaign, 2017

Urbana, Illinois

Doctoral Committee:

Assistant Professor Jana Diesner, Chair
Associate Professor Catherine L. Blake
Assistant Professor Vetle I. Torvik
Associate Professor Michelle Shumate, Northwestern University
Dr. Seok-Hyoung Lee, Korea Institute of Science and Technology Information

ABSTRACT

In this study, I demonstrate that the choice of disambiguation methods for resolving author name ambiguity can adversely affect our understanding of scholarly collaboration patterns and coauthorship network structures extracted from large-scale scholarly data. By utilizing large-scale bibliometric data, scholars in many fields have gleaned knowledge for use in scholarly evaluation, collaborator recommendations, research policy evaluation, and network-evolution modeling. A common challenge has been that author names in bibliometric data are not properly disambiguated: authors may share the same name (i.e., different authors are sometimes misrepresented to be a single author which can lead to a "merging of identities"). In addition, one author may use name variations (i.e., an author may be represented as two or more different authors which can lead to a "splitting of identities"). When faced with these challenges, most scholars have pre-processed bibliometric data using simple heuristics (e.g., if two author names share the same surname and given name initials, they are presumed to represent the same author identity) and assumed that their findings are robust to errors due to author name ambiguity. I test this long-held assumption in bibliometrics by measuring the impact of author name ambiguity on network properties. I accomplish this under varying conditions, including network size and cumulative time window (from 1991 to 2009) using four large-scale bibliometric datasets that cover: biomedicine, computer science, psychology and neuroscience, and one nation's entire domestic publication output. For this task, I collate the statistical properties of coauthorship networks constructed from algorithmically disambiguated data (i.e., close to clean data) against those that come from the same networks, but are compromised by misidentified authors via first-initial and all-initials disambiguation methods. In addition, I simulate the levels of merging and splitting incrementally using those empirical datasets. My findings show that initial-based name disambiguation methods can severely distort our understanding of given networks and such distortion gets worse over time. Moreover, the distortion sometimes leads to biased or false knowledge of coauthorship network formation and evolution mechanisms such as preferential attachment generating the power-law distribution of vertex degree and to false validation of theories about the choice of collaborators in scientific research. This may result in ill-informed decisions about research policy and resource allocation. Besides measuring the impact of name ambiguity on network properties, I also test how name ambiguity can be estimated using simple heuristics such as dataset size and how merged author identities can be detected via an author's ego-network properties to provide a practical guidance for corrective measures. My research calls for further studying the effects of author name ambiguity on coauthorship network properties and is expected to help scholars establish better practices for knowledge discovery from large-scale scholarly data.

ACKNOWLEDGEMENTS

This thesis was supported by the grants from the Korea Institute of Science and Technology Information (2014~2016), the Beta-Phi-Mu Eugene Garfield Doctoral Dissertation Fellowship (2016~2017), Microsoft Azure for Research Award (2016~2017), and the Elsessers Fellowship (2017). I would like to thank my committee members and family for support and help. Above all, I would like to thank my advisor, Dr. Jana Diesner, who has inspired and guided me.

TABLE OF CONTENTS

CHAPTER 1: INTRODUCTION

1.1 MOTIVATION

This thesis is motivated by the observation that scholars in biomedicine, chemistry, computer science, and biomedicine, among other fields, have developed and evaluated advanced network metrics and algorithms for coauthor recommendation, community detection, network evolution, and vertex ranking based on various real-world coauthorship networks where author names are not properly disambiguated. Those scholars have based their studies on the assumption that their network data are error-free or that their findings are robust to data errors due to author name ambiguity. Recently, some studies began to test this assumption. For this purpose, they compare properties of coauthorship networks constructed from close-to-clean data (i.e., with algorithmically disambiguated author names) versus the same networks but compromised by errors in merged and/or split author identities. Their findings showed that the quality of network data can severely distort both our micro- (e.g., vertex ranking) and macro-level (e.g., topology) understanding of a given coauthorship network, sometimes leading to false positive findings of network evolution mechanisms (e.g., power-law distribution of vertex degree) or of network coherence. This thesis[1] aims to measure the impact of author name ambiguity on network properties under varying conditions such as network size and time window in order to test whether compromised author names can be identified by their network-based characteristics, and to provide practical guidance for scholars and practitioners to help them improve decision making based on compromised coauthorship networks.

[1] I rely heavily in this chapter on Kim and Diesner (2015), Kim and Diesner (2016), Kim, Kim, and Diesner (2014). Especially, Kim, Kim, and Diesner (2014) has been published under the terms of the Creative Commons Attribution License (http:/ creativecommons.org/licenses/by/3.0/).

1.2 KNOWLEDGE DISCOVERY FROM NETWORK DATA

Scholars have used large-scale bibliometric data to understand the global structure of scientific collaboration in various fields. A well-known example is Newman (2001b, 2004), which studied coauthorship networks in biology (1,520,251 vertices), physics (52,909 vertices), computer science (11,994 vertices), and mathematics (253,339 vertices). The studies showed that biology scholars produce more papers and have more coauthors than scholars in the other academic domains. Furthermore, the studies argued that the coauthorship networks produced a power-law distribution of vertex degree. Another exemplar work is Barabási et al. (2002), in which coauthorship networks in mathematics (70,975 vertices) and neuroscience (209,293 vertices) were also found to produce a power-law distribution of vertex degree. A noticeable argument of the study was that preferential attachment, a tendency of newcomers to the network to attach to vertices with a high degree, could produce the power-law distribution. These studies have attracted significant scholarly attention and researchers in diverse fields have conducted benchmark studies confirming, among other findings, the power-law distribution of vertex degree and a Small-Worldness of coauthorship networks (e.g., Börner, Maru, & Goldstone, 2004; Milojević, 2010; Perc, 2010). Such law-abiding characteristics of coauthorship networks also motivated several scholars to model international collaboration as a self-organizing complex system following a preferential attachment mechanism and propose that the mechanism can be used to create the "most efficient organization of researchers" across the world (e.g., Wagner, 2009, p.108).

The findings from the afore-mentioned studies were obtained through network analysis. A network is defined as a set (group) of vertices connected by edges, which represent relationships between vertices such as communication, friendship, or flow of electricity (Newman, 2010).

From the metadata of publications such as title, author names, and years, scholars construct coauthorship networks for a snapshot or longitudinal view of scientific collaboration. In coauthorship networks, vertices represent authors who are linked if they have collaborated together on a paper. This conceptualization involves the projection of two-mode (a.k.a. bipartite network, where authors are connected to papers they wrote) into one-mode (a.k.a. monopartite network, where only authors are connected by co-appearance in the bylines of papers). Technically, network construction from bibliometric data begins with identifying vertices and edges from network data.

> (1) Usually, vertices (authors) in coauthorship networks are identified by name strings (i.e., author names) in the byline of a paper. Bibliometric data used for coauthorship network construction are semi-structured: a publication record has information clearly defined by data publishers such as paper title, author list (i.e., byline), publication venue, publication year, or abstract. Thus, author name detection itself is not a problem because author names are clearly expressed in the author byline.
>
> (2) Next, edges are formed between those identified vertices (authors) that have ever appeared together in the author list of a paper. This step corresponds to Relation Extraction. Identifying edges is not a problem in coauthorship network construction because the co-appearance of author names in a byline clearly indicate the existence of coauthoring relationship (edges) between authors.

1.3 THE CHALLENGE OF AUTHOR NAME AMBIGUITY RESOLUTION

A challenge in constructing coauthorship networks arises when one or more author names may refer to the same identity (Diesner, 2012; Diesner & Carley, 2009). For example, a scholar can be represented by different author names due to inconsistent spelling of middle names or recording errors. For example, the same person may have two name variants, “John Doe” in one paper and “John M. Doe” in another paper. Here, a single identity is represented by two different author names: the splitting of an identity. Another challenging situation is the merging of identity: two different author identities can be represented by the same author name because they happen to have the same name. For example, “Mark Newman,” a physicist at the University of Michigan can be regarded as the same person as “Mark Newman,” a communication scholar at the University of Michigan. Splitting or merging of identities requires ambiguity resolution.

In computer and information science, ambiguity resolution of author names has been actively studied under the names of record linkage, deduplication, co-reference resolution, and authority control (e.g., Bhattacharya & Getoor, 2005; Culotta & McCallum, 2005; Sarawagi, 2008). Findings from such studies have been, however, rarely applied to coauthorship network studies with a few exceptions (e.g., Fegley & Torvik, 2013; Strotmann, Zhao, & Bubela, 2009). Meanwhile, ambiguity of author names in bibliometric network data has not yet received proper attention from scholars who use bibliometric data for research.

> (1) The majority of coauthorship network papers does not discuss the problem of author name ambiguity (Kim, Kim, et al., 2014). Some studies acknowledge the problem but do not resolve name ambiguity on purpose because, for example, it may introduce noise into the data (e.g., Braun, Glänzel, & Schubert, 2001; Larivière, Sugimoto, & Cronin, 2012; Wagner & Leydesdorff, 2005).
>
> (2) A small number of papers disambiguated author names manually or computationally by using author affiliation information or CVs available online (e.g., Chua & Yang, 2008; Strotmann et al., 2009; E. J. Yan & Ding, 2009).

Between the two extremes – no disambiguation and disambiguation through additional author information – lies the initial-based disambiguation method. This approach relies on the initials of a given name(s) of an author. Two variations have been widely used:

> (1) First-initial method: several studies assume that, if two author names share a full surname (or last name) and the first initial of a given name, they represent the same author identity (e.g., Liben-Nowell & Kleinberg, 2007). According to this method, "Newman, Mark E. J." and "Newman, Mark" refer to the same identity because they both have "M" in the given names, even if they represent different identities.
>
> (2) All-initials method: others rely on all initials of given names (e.g., Milojević, 2010). This method regards "Newman, Mark E. J." and "Newman, Mark" as referring to different identities because they don't share middle name initials. This scheme, however, ignores the possibility that the two author names are name variations of a single author.

These simple, heuristic disambiguation techniques have been widely used by coauthorship network researchers (Milojević, 2013; Strotmann & Zhao, 2012). Table 1 illustrates some selected large-scale coauthorship network studies that use initial-based disambiguation[2]. Most of these example studies acknowledge that initial-based disambiguation may induce misidentification errors, but argue that the effects of misidentification on research findings are negligible (e.g., Barabási et al., 2002; Liben-Nowell & Kleinberg, 2007; Milojević, 2010; Newman, 2001, 2004).

[2] To select the papers, the author searched journal papers with 'network' or 'networks' in titles indexed by ISI Web of Science Core Collection. The output list was filtered for the top 200 papers by citation counts. Then, papers that analyze coauthorship networks with at least 10,000 vertices were selected. Among them, eight papers were finally selected considering academic fields.

Table 1: Examples of Coauthorship Network Studies (Reused Table 1 from Kim and Diesner (2016))

Field	**Research**	**Data Source**	**Year (Period)**	**No. of Journals**	**No. of Articles**	**No. of Estimated Vertices**	**Disambiguation Method**
Biomedicine	Newman (2001, 2004)	MEDLINE	1995-1999 (5)	Not Specified	2,163,923	1,520,251	First-Initial All-Initials
Computer science	Fiala (2012)	Web of Science	1996-2005 (10)	426	205,780	187,016	All-Initials
Nanoscience	Milojević (2010)	NanoBank	2000-2004 (5)	4,792	270,135	294,456	All-Initials
Neuroscience	Barabási et al. (2002)	Unknown	1991-1998 (8)	“all relevant journals”	210,750	209,293	All-Initials
Physics	Radicchi et al. (2009)	American Physics Society	1893-2006 (114)	Physical Review Collection	407,236	216,623	All-Initials
	Liben-Nowell et al. (2007)	arXiv	1994-1999 (6)	Subfields in Physics	35,555	23,589	First-Initial
Inter-disciplinary	Börner et al. (2004)	Web of Science	1982-2001 (20)	1 (PNAS)	45,120	105,915	All-Initials
	Petersen et al. (2011)	Web of Science	1958-2008 (51)	6	311,880	634,288	All-Initials

1.4 MEASURING THE IMPACT OF AUTHOR NAME AMBIGUITY

Recently, a few scholars began to question the supposedly negligible impact of author name misidentification on research findings. They showed that coauthorship network properties can be changed by merged or split author identities via initial-based disambiguation (Diesner et al., 2015; Fegley & Torvik, 2013; Kim & Diesner, 2015, 2016; Kim, Diesner, et al., 2014; Kim, Kim, et al., 2014). For example, initial-based disambiguation was found to decrease the number of authors (i.e., vertices), average shortest path lengths, degree assortativity, transitivity, and number of network components, while it increases average production of authors, average degree, network density, and size of the largest components (Fegley & Torvik, 2013; Kim & Diesner, 2015).

1.5 AIMS AND CONTRIBUTIONS

Following this line of research, this thesis aims to obtain a deeper understanding of the percussion of author name disambiguation on the structure and evolution of large-scale coauthorship networks. In particular, this thesis attempts to address the following challenges that have been insufficiently dealt with in previous studies.

(1) Measuring the impact of errors in author name disambiguation in large-scale coauthorship networks: with different levels of disambiguation errors, network size, and time window.

(2) Testing whether such errors can be estimated by network-based characteristics of authors in coauthorship networks.

(3) Based on the analysis in (1) and (2), providing suggestions to improve name disambiguation efforts and decision making based on ambiguous data.

Considering the dominant practice of ignoring the need for author name disambiguation (Kim, Kim, et al., 2014) and the frequent use of initial-based disambiguation in bibliometrics (Zhao & Strotmann, 2011), investigating the impact of author name disambiguation on our understanding of network properties is of great importance to the scholarly community. Specifically, as coauthorship networks have been used to test hypotheses and answer theoretical and empirical research questions about networks, a deeper understanding of the effects of author name ambiguity can “contribute to a greater comparability and generalizability of findings” from both previous and future research (Diesner, Evans, & Kim, 2015). The obtained knowledge on network-based characteristics of ambiguous names and its mechanism of distorting network properties can help scholars gain proper insights from increasing network studies.

1.6 ORGANIZATION OF STUDY

This study is organized as follows. In the Chapter 2, how disambiguation errors produced by ambiguous author names have been discussed in prior research is reviewed. Next, a description of four datasets for analysis is provided. After that, a list of measurements used in the thesis is defined, followed by introducing results from the analysis. In addition, results of estimating error levels in publication records and detecting compromised authors are reported. Finally, contributions, real-world implications, and limitations are discussed.

CHAPTER 2: LITERATURE REVIEW

In this chapter, I review how scholars have addressed the resolution of author name ambiguity in bibliometric data[3].

2.1 MANUAL IDENTIFICATION OF AUTHOR NAMES

Facing the need to resolve the ambiguity of author names, several scholars manually checked the identity of each author name. For example, ambiguous names of authors in a dataset were compared by affiliation information associated with each name (e.g., Chua & Yang, 2008; E. J. Yan & Ding, 2009). If two ambiguous author names shared the similar or same affiliation, they were regarded to refer to the same identity. During this process, researchers considered additional information sources such as the scholar's personal or institutional webpage, CVs available online, or they sent correspondence to authors with ambiguous names to confirm whether a specific paper was written by them.

This approach is not scalable for large-scale data and can be costly. A more critical problem is that this manual inspection does not "guarantee perfect disambiguation" (Torvik & Smalheiser, 2009). This is because additional information may not be available for some or many of the ambiguous author names, which may leave the decision of a name's identity difficult. Another issue is that studies relying on this method rarely report the accuracy of their ambiguity resolution efforts nor the ratio of agreement on matched or unmatched cases between two or more coders (e.g., Acedo, Barroso, Casanueva, & Galan, 2006; Chua & Yang, 2008; E. J. Yan & Ding, 2009; Yoshikane, Nozawa, Shibui, & Suzuki, 2009). Moreover, these studies did not

[3] I rely heavily in this chapter on Kim and Diesner (2015), Kim and Diesner (2016), Kim, Kim, and Diesner (2014), and Kim, Diesner, Kim, Aleyasen, and Kim (2014).

consider the comparison of research findings obtained from before- to after-disambiguated network data.

2.2 ALGORITHMIC DISAMBIGUATION

Resolving author name ambiguity has been actively studied by computer and information scientists in the areas of Natural Language Processing and Information Extraction (Sarawagi, 2008). To obtain high accuracy in disambiguating author names, researchers have typically taken two steps (Treeratpituk & Giles, 2009).

(1) First, author names in bibliometrics data are compared for similarity on attributes such as coauthor name, affiliation, paper title, venue, or citing references. This procedure usually produces a binary decisions (matched or unmatched) or a similarity score (usually between 0 and 1).

(2) Next, based on the similarity profile, author name instances are put into clusters if they are decided to refer to the same identity. Here, each cluster represents a unique identity.

Although some papers deal with both similarity comparison (Step 1) and clustering (Step 2), most algorithmic disambiguation studies have focused on the first step because finding pairwise linkage function that provides high accuracy can lead to a high-quality clustering outcome. To find the best performing function, various supervised and unsupervised machine learning algorithms have been used (For a review on this, refer to Ferreira, Gonçalves, & Laender, 2012). To test the performance of name disambiguation, a sample of clusters (usually the most difficult cases such as "Wang, J.") are manually checked (as described in 2.1 Manual Identification of Author names) to generate ground-truth data, and the performance of an algorithm is measured against the ground-truth by evaluation metrics such as pairwise F1, K-metric, or cluster F1.

Most algorithmic disambiguation studies utilized various features extracted from meta-data or text data such as affiliation, email address, paper title, keyword, publication venue name, and abstract (Ferreira et al., 2012). However, such meta- and text data may be imperfect or unavailable. Let's take affiliation for example. It is possible to see that (1) authors change academic affiliations several times during their academic career, (2) affiliation information is missing or not matched with authors, and (3) an affiliation has name variants due to official name changes or inconsistent records by authors. Especially for Case (3), affiliation name disambiguation may need to be conducted as a prerequisite for author name disambiguation (e.g., Deville et al., 2014; Martin, Ball, Karrer, & Newman, 2013). All of these factors can affect the performance of similarity measures in algorithmic author name disambiguation.

Another limitation of algorithm-based approaches is that most studies are confined to specific fields, especially computer science, in terms of their application domain. This domain-specificity can limit the applicability and generalizability of algorithmic disambiguation to other domains. For example, features extracted from coauthorship or titles in computer science data can lead to lower performances when applied to chemistry, where large team-based collaboration is dominant, and many title words include chemical symbols and a mixture of alphanumeric and special characters. The most important issue with algorithmic disambiguation studies is that they do not utilize disambiguated data for network analysis, although a few exceptions exist (e.g., Deville et al., 2014; Fegley & Torvik, 2013; Martin et al., 2013).

2.3 INITIAL-BASED DISAMBIGUATION

In studies using scholarly data, initial-based author name disambiguation is the dominant mode of handling the ambiguity of author names (Milojević, 2013; Strotmann & Zhao, 2012). Such a pratice is partly because most author names in bibliometric data used in previous studies were recorded in the "a full surname plus a given name initial(s)" format. Three initial-based disambiguation methods have been used in practice.

(1) First-initial Method: if two author names share the first initial of their given names, they represent the same author identity (e.g., Bettencourt, Kaiser, & Kaur, 2009; Ding, 2011; Goyal, van der Leij, & Moraga-Gonzalez, 2006; Liben-Nowell & Kleinberg, 2007). Since this method does not consider situations where initials of other given names exist or are different, it can produce a disambiguation error by representing two different author names as a single identity (i.e., merging). For example, "Blake, C. L." can be fused with "Blake, C. C." into "Blake, C." although they may not present the same person.

(2) All-initials Method: author name instances that share all the initials of first and middle (if any) given names belong to the same author identity (e.g., Barabási et al., 2002; Fiala, 2012; Milojević, 2010; Newman, 2001b, 2004; Radicchi, Fortunato, Markines, & Vespignani, 2009; Rorissa & Yuan, 2012). For example, "Blake, C. L." and "Blake, C. C." would refer to different author identities. This disambiguation method may produce splitting errors: e.g., "Blake, C. L." is different from "Blake, C.," even though these two name instances may refer to the same identity. Here, the author may omit the second initial in one paper and not in another.

(3) Hybrid Method: If an author name has a first given name initial and can be matched with two or more names with different secondary given name initials, all these name instances relate to different author identities (e.g., Milojević, 2013; Yoshikane et al., 2009). For example, if "Blake, C." is compared with "Blake, C. L." and "Blake, C. C.," these three author names are thought to represent three different author identities. Here, all-initials disambiguation applies. If "Blake, C." has one candidate name to match, "Blake, C. L.,"

these two name instances represent the same author. Here, first-initial disambiguation applies. Thus, this approach is called a hybrid method, i.e., a mixture of the first-initial and all-initials methods. As this method combines (1) and (2), it can have either merging or splitting errors, or sometimes both types of errors.

Once processed by one of these three methods, author names that are supposed to represent the same identity are grouped into a cluster. This procedure is the same as that described in the previous section, 2.2 Algorithmic Disambiguation. The difference lies in that the inter-name similarity is measured only by matching a full surname and given name initials. In this sense, initial-based disambiguation may be called an algorithmic disambiguation in a simplified form.

Many scholars relying on initial-based disambiguation have agreed that their method can produce errors in disambiguating author names. The problem is, however, that most of them have not attempted to measure or report these errors. Instead, some scholars just took extra disambiguation steps using affiliation information to reduce such errors (e.g., Yoshikane et al., 2009), but without reporting how such measures improve accuracy.

Interestingly, Newman (2001) proposed that the numbers of authors disambiguated by the first-initial and all-initials methods represent “the lower and upper bounds” of the “true” number of unique author identities (i.e., first-initial → lower bound & all-initials → upper bound). Then, he calculated properties of several coauthorship networks pre-processed by the proposed first-initial and all-initials approaches, and reached the conclusion that most network properties produced errors of “an order of a few percent” between these two networks. Based on these findings, he argued that the properties of “true” networks can be found between these two extremes.

Citing Newman (2001), many scholars chose initial-based name disambiguation for disambiguating author names and justified their choice stating that they “believe” or “assume” that the errors in identifying authors do not have much impact on research findings (e.g., Barabási et al., 2002; Goyal et al., 2006; Liben-Nowell & Kleinberg, 2007; Milojević, 2010; Yoshikane et al., 2009).

2.4 ‘NO DISAMBIGUATION’ APPROACH

The most prevalent approach towards author name ambiguity in bibliometrics seems to ignore or not comment on the issue (Kim, Kim, et al., 2014). This means that author names in bibliometric data are assumed to represent unique author identities. Some users of this approach clearly indicate that no author name disambiguation was performed because algorithmic disambiguation does not “guarantee perfect disambiguation” and, sometimes, may introduce noise into the data (e.g., Börner et al., 2004; Wagner & Leydesdorff, 2005).

This approach can be the same as either first-initial or all-initials method when names recorded in the scholarly data are documented in the record scheme of a full last name (surname) and an initial(s) of a given name(s). This may happen frequently because many bibliometric studies obtain data from a few representative scholarly data services, where author names are provided in the record format of a last name followed by an initial(s) of a given name(s).

2.5 IMPACT OF NAME DISAMBIGUATION ON NETWORK PROPERTIES

As described above, studies using manual or algorithmic name disambiguation usually have not paid attention to its impact on network properties. In contrast, network researchers who apply initial-based methods to disambiguating names have recognized the possibility of disambiguation errors affecting network properties. The issue is, however, that “the assumption of a supposedly negligible effect of name disambiguation errors in large-scale coauthorship networks has not been tested in a rigorous way" (Kim & Diesner, 2016).

A notable exception to this tendency is Milojević (2013). The study tested how well initial-based author name disambiguation performs using synthetic data. For this purpose, specifically, the synthetic data were simulated as ground-truth using the frequencies of surnames and given name initials from real-world scholarly datasets representing several academic fields. The main finding was that the ratios of estimated "true" identities that were contaminated by first-initial, all-initials, and hybrid methods were 1.5% up to 5.5% (if only best performance was reported). Its conclusion was that, if used "within a single discipline or a field that use only information contained in names," initial-based author name disambiguation can be "quite accurate" and is supposed not to "have an adverse effect on many or most statistical bibliometrics studies" (Milojević, 2013, p. 773). The simulation study, however, did not consider the errors in the synthetic data.

Recently, several scholars have scrutinized the accuracy of initial-based author name disambiguation. Fegley and Torvik (2013) generated a large-scale coauthorship network from a ground-truth dataset (where author names are disambiguated with advanced algorithms described in Torvik and Smalheiser (2009)), and another network from the same dataset but with author names disambiguated by first-initial and all-initials. The authors showed that initial-based disambiguation "dramatically" inflates or deflates coauthorship network properties. For example, the number of vertices (i.e., unique authors) identified in the algorithmically disambiguated network was reduced from 3.17 to a) 1.56 million identities by the first-initial disambiguation and b) 2.18 million by the all-initials disambiguation (Torvik & Smalheiser, 2009). Another group of scholars extended Fegley and Torvik's work (2013) by describing the distortive effects of initial-based disambiguation on bibliometric data from diverse fields such as computer science, biology, nanoscience, and mathematics, and at a national level, i.e., domestic coauthorship networks in Korea (Diesner et al., 2015; Kim & Diesner, 2016; Kim, Diesner, et al., 2014; Kim, Kim, et al., 2014).

Those studies, however, did not address several issues. First, it is not clear how name disambiguation errors (i.e., merging and splitting) can affect network properties over time in other data. Second, the relationship between the disambiguation errors and dataset size has not been studied. For example, as the size of scholarly datasets becomes smaller, the disambiguation errors produced by initial-based methods may also decrease to an extent that will not change or distort research findings. Third, it has not been sufficiently discussed what type of network properties are relatively immune or vulnerable to name disambiguation errors and under what levels of name disambiguation errors or accuracy. Fourth, estimation of name disambiguation levels and detection of compromised (merged or split) author identities have not been researched.

2.6 RESEARCH QUESTIONS

This study addresses the afore-mentioned issues by seeking answers to the following questions:

(1) To what extent are network metrics and topologies (power-law distribution) affected by compromised vertices and/or edges in large-scale coauthorship networks constructed from various data? How does such impact change over time?

(2) What levels of name disambiguation errors are associated with how much distortion of network measures? What are the acceptable levels of disambiguation errors that can reduce the distortive impact to a negligible extent?

(3) Are levels of disambiguation errors predictable? If so, what predictors perform best?

(4) What are the network-based characteristics of compromised author names? What are the best methods to detect compromised author names and eliminate (or remedy) their impact on network properties?

(5) What do the findings from (1) ~ (4) suggest for researchers and practitioners who analyze error-prone data?

Why should we care about these questions? Findings from contaminated coauthorship networks can mislead us to a flawed understanding of the structure of scientific collaboration, invalid inferences about its underlying mechanisms, and, thus, affect hypotheses testing, theory building, and decision making for academic resource allocation in a field or at a national level. The research questions above are expected to provide us a better knowledge of the impact of author name disambiguation on network properties and to help us draw correct implications from coauthorship network studies.

In addition, answers to these questions will help us decide whether we could ease concerns about disambiguation errors or if we should be cautious about even minor changes of disambiguation accuracy in network data. In other words, this research can provide a warning or a go-signal to research. For example, as large-scale bibliometric data are being accumulated at an unprecedented volume and rate coupled with powerful computational capacity at hand, scholars now have an opportunity to study scholarly communication at a scale which previous scholars could not attempt to do. With better knowledge of the relationship between disambiguation errors and their impact on network of varying size, we could advise scholars and practitioners on why author name disambiguation in scholarly data matters and is worth the costs for data quality control. In the following two chapters, datasets and measurements used in this thesis for answering these research questions are detailed.

CHAPTER 3: DATA

To address the outlined research questions, this dissertation uses various real-world bibliometric data disambiguated by algorithms[4].

3.1 MEDLINE

MEDLINE refers to the bibliographic database which is maintained by the National Library of Medicine and covers publications in medicine research. The dataset is released to the public in XML format. Each paper in the data is recorded with a unique identifier (PMID), paper title, journal title, author names, author affiliations (if available), and medical subject headings (MeSH), which are predefined topic categories assigned manually by human experts.

Author names in MEDLINE are not disambiguated. Disambiguated author names were retrieved from Author-ity (Torvik & Smalheiser, 2009) data, which contain MEDLINE author names disambiguated with an accuracy of 98~99% via advanced algorithms and statistical modeling. According to Torvik and Smalheiser (2009), pairs of author names were selected by name string matching rules. Then, they were compared for calculating similarity based on features extracted from papers' metadata: initial of a middle name(s), name suffix (e.g., Jr. or II), journal title, coauthor names, words in paper title, words in affiliation name, language used in paper, and MeSH term. When the "combination of match values from these eight features passed a certain threshold value, the target name pairs were merged via a maximum likelihood based, agglomerative algorithm" (Diesner, Evans, & Kim, 2015; for details, refer to Torvik & Smalheiser, 2009 and Torvik, Weeber, Swanson, & Smalheiser, 2005).

[4] I rely heavily in this chapter on Kim and Diesner (2015), Kim and Diesner (2016).

For this thesis, a subset of Author-ity data was used. A total of 1,551,483 papers that have been published between 1991 and 2009 with the MeSH term "Physiology" were chosen for analysis.

3.2 DBLP

The Digital Bibliography & Library Project (DBLP) data index metadata of books, conference proceedings, and journal publications in computer science and its related fields such as discrete mathematics and informatics. DBLP data are freely available in XML format and have been used by scholars to study patterns of collaboration and to model network evolution (e.g., Biryukov & Dong, 2010; Franceschet, 2011). Author names in the DBLP data are disambiguated by algorithms and manual inspection. First, they are disambiguated by heuristic rules for matching text strings of author names, followed by similarity calculation based on matching coauthors of an author and coauthors of coauthors of the author. Then, DBLP accepts input from scholars who want to correct their bibliometric entry, which is believed to contribute to the accuracy of name disambiguation in DBLP. These two steps of disambiguation are repeated regularly (Reitz & Hoffmann, 2010).

Although scholars have argued that DBLP is "internationally respected" for its accuracy in name disambiguation (Franceschet, 2011), its accuracy has rarely been tested. This thesis followed the method described in Kim and Diesner (2015) to test the performance of DBLP's author name disambiguation using a ground-truth dataset of 476 unique authors in 6,517 publications who have ambiguous surnames such as 'Kim' or 'Johnson.' The ground-truth dataset was generated originally by Han, Zha, and Giles (2005), but corrected for errors by Shin, Kim, Choi, and Kim (2014). During the process of matching paper records in the ground-truth data with those in DBLP, a total of 3,921 papers (474 unique authors) were found to match. The mismatch is because publications in the ground-truth dataset include papers published in journals or books that are not indexed by DBLP.

Two metrics, K-metric and pairwise F1, were used to measure the performance of DBLP's disambiguation. These metrics have been frequently used for measuring the performance of name disambiguation algorithms in computer science (Ferreira et al., 2012).

K-metric: "This is the geometric mean of average cluster purity (ACP) and average author purity (AAP)[5].

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

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

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

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*" (Kim & Diesner, 2015).

"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

[5] The paragraphs (in quotation marks) describing K-metric and Pairwise F1 were re-used from Kim and Diesner (2015)

this cluster but are not included in it (low splitting), the AAP value gets closer to 1" (Kim & Diesner, 2015).

Pairwise F1: "Pairwise precision (*p*P) is calculated as *p*P = A/(A+C), while pairwise recall (*p*R) is calculated as *p*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 *p*F1 is defined as follows:

$$pF1 = \frac{\left((\beta^2 + 1) \times pP \times pR\right)}{(\beta^2 \times pP + pR)}$$

Here, β is the weight of recall relative to precision. We use $\beta = 1$ as in F1, which weighs two metrics equally" (Kim & Diesner, 2015).

Results of the disambiguation tests are summarized in Table 2. The DBLP accuracy for the May 2014 version (which was used for this thesis) was on average 0.952 in terms of K-metric and 0.96 in terms of Pairwise F1. This decent performance is comparable to that of other disambiguation algorithms which showed similar or slightly lower scores (Cota, Ferreira, Nascimento, Goncalves, & Laender, 2010; Ferreira et al., 2012; Pereira et al., 2009).

Table 2: Performance Evaluation of Name Disambiguation in DBLP

Name String	Name Instances	Unique Identities	K-Metric			Pairwise F1		
			ACP	AAP	K-Value	Recall	Precision	F1
A. Gupta	370	26	0.963	0.979	0.966	0.98	0.98	0.98
A. Kumar	145	14	0.953	0.974	0.964	0.97	0.98	0.97
C. Chen	436	61	0.936	0.979	0.957	0.98	0.94	0.96
D. Johnson	178	15	0.932	0.959	0.946	0.97	0.98	0.98
J. Lee	664	99	0.912	0.965	0.938	0.96	0.92	0.94
J. Martin	83	15	0.964	0.982	0.973	0.99	0.97	0.98
J. Robinson	105	12	0.970	0.977	0.973	0.99	0.98	0.99
J. Smith	336	29	0.933	0.985	0.959	0.99	0.97	0.98
K. Tanaka	144	10	0.990	0.988	0.989	1.00	1.00	1.00
M. Brown	79	13	0.901	0.956	0.928	0.95	0.88	0.92
M. Jones	110	13	1.000	0.950	0.975	0.94	1.00	0.97
M. Miller	92	12	0.986	0.965	0.975	0.97	1.00	0.98
S. Lee	741	84	0.914	0.969	0.941	0.97	0.91	0.94
Y. Chen	438	71	0.898	0.987	0.941	0.99	0.86	0.92
Total or Avg.	3,921	474	0.947	0.973	0.959	0.98	0.96	0.97

For this thesis, a total of 1.397,870 papers published between 1991 and 2009 in all the conference proceedings and journals indexed in DBLP were selected.

3.3 MAG

Microsoft Academic Graph (MAG) is a bibliometric data service provided by Microsoft. MAG records metadata of more than 90 million publications in conference proceedings and journals across all scholarly domains. The baseline data that are used for the service are freely available for research purposes. The MAG management team argues that author names in the data are disambiguated by using “various best-effort algorithms” without further detailed explanation (e.g., feature selection) on those algorithms or any information on accuracy of disambiguation (Sinha et al., 2015).

This thesis tested the accuracy of name disambiguation in MAG against the same ground-truth data used for DBLP. Table 3 reports the results. The accuracy of name disambiguation in MAG is lower than that of DBLP: on average, 0.772 for K-metric and 0.72 for pairwise F1. The levels of accuracy are decent when compared to other disambiguation studies. A noticeable observation from the table is that MAG’s disambiguation algorithms performed well for reducing merged identities: i.e., high scores on ACP in K-metric and Precision in pairwise F1. In contrast, the low scores on AAP in K-metric and Recall in pairwise F1 indicate that the name disambiguation algorithm in MAG was vulnerable to splitting. For the analysis of this thesis, a total of 573,816 papers published between 1991 and 2009 in journals indexed as Psychology for the journal category was selected. This dataset contains approximately 300,000 papers in Neuroscience corresponding to the data used in Barabási et al. (2002).

Table 3: Evaluation of Name Disambiguation in MAG

Name String	**Name Instances**	**Unique Identities**	**K-Metric**			**Pairwise F1**		
			ACP	**AAP**	**K-Value**	**Recall**	**Precision**	**F1**
A. Gupta	91	13	0.980	0.659	0.803	0.68	0.97	0.80
A. Kumar	453	61	0.947	0.648	0.783	0.57	0.94	0.71
C. Chen	623	71	0.927	0.585	0.736	0.37	0.92	0.52
D. Johnson	382	26	0.895	0.478	0.654	0.33	0.91	0.48
J. Lee	218	15	0.941	0.526	0.703	0.47	0.95	0.63
J. Martin	148	13	0.957	0.658	0.793	0.66	0.97	0.79
J. Robinson	145	14	0.967	0.458	0.665	0.35	0.98	0.52
J. Smith	769	99	0.947	0.595	0.751	0.47	0.96	0.63
K. Tanaka	780	84	0.932	0.670	0.790	0.68	0.94	0.79
M. Brown	74	15	0.986	0.683	0.821	0.68	0.99	0.81
M. Jones	187	12	0.974	0.750	0.855	0.91	0.96	0.93
M. Miller	113	12	0.991	0.594	0.767	0.55	1.00	0.71
S. Lee	509	30	0.902	0.745	0.820	0.83	0.90	0.87
Y. Chen	183	10	0.992	0.756	0.866	0.77	1.00	0.87
Total or Avg.	4,675	475	0.953	0.629	0.772	0.59	0.96	0.72

3.4 KISTI

The Korea Institute of Science and Technology Information (KISTI) is an organization in control of collecting, processing, analyzing, and disseminating the information of research publications published in South Korea. KISTI used a two-stage name disambiguation process. At the first stage, “author name instances were clustered algorithmically using features such as author name string, affiliation, coauthor, keywords of the publication, and publication venue” (Kim, Tao, Lee, & Diesner, 2016). After this stage, the accuracy of author name disambiguation in KISTI data is known to be 0.94 (pairwise F1) against a sample of more than 30,000 names. Then, human experts investigate clusters that are believed to be erroneous and correct the errors of those clusters manually, which increase the overall accuracy up to 98% (Kim et al., 2016). The KISTI data records most author names in English. Some of them were transliterated into English from Korean or Chinese by KISTI. For this thesis, a total of 507,399 papers published between 1991 and 2009 in conference proceedings and journals were filtered. This selection corresponds to the dataset used in Çavuşoğlu and Türker (2013).

CHAPTER 4: METHODOLOGY

4.1 GENERAL STRATEGY

One way to measure the impact of name disambiguation methods on coauthorship network properties is to note the similarity of the properties calculated for two or more coauthorship networks: one constructed from disambiguated data and others constructed from the same data that has not been disambiguated or has been disambiguated by different methods[6]. Here, the algorithmically disambiguated data serve as a *proxy* of ground-truth. This idea was introduced by social network scientists who tested how stable and robust network measures are when network data are contaminated by the addition or removal of vertices and edges (e.g., Borgatti, Carley, & Krackhardt, 2006; Diesner & Carley, 2009; Frantz, Cataldo, & Carley, 2009).

To investigate the effects of author name disambiguation on network properties, three coauthorship networks were generated for each dataset: (1) a network from disambiguated data, (2) a network that was constructed from the same dataset with author names in it pre-processed by the first-initial disambiguation method, and (3) a network that was constructed from the same dataset with author names in it pre-processed by the all-initials disambiguation method.

In addition to a snapshot view of each network, longitudinal changes in network properties are traced with a yearly resolution. For this task, accumulative networks up to a target year (ranging from 1991 to 2009, 19 years) were created for each of the three types of networks per each dataset. This cumulative time slicing has been used in previous network studies to investigate

[6] I rely heavily in this chapter on Kim and Diesner (2015), Kim and Diesner (2016), Kim and Diesner (2017), Kim, Kim, and Diesner (2014), and Kim, Diesner, Kim, Aleyasen, and Kim (2014).

network evolution and test edge formation mechanisms (e.g., Barabási et al., 2002; Çavuşoğlu & Türker, 2013; Franceschet, 2011; Kim & Diesner, 2015; Kim et al., 2016; Perc, 2010).

Algorithmically disambiguated networks are used as *proxies* of ground-truth data against which other types of networks (i.e., where initial-based disambiguation is performed) are compared to see the changes in network properties.

To disambiguate author names based on given name initials, each author name string in the data (e.g., Blake, Catherine L.) was changed into two variations: (1) full surname, a comma, and the first initial of a given name (s) (e.g., Blake, C.) and (2) full surname, a comma, and all initials of a given name(s) (e.g., Blake, C. L.). The Hybrid method of initial-based disambiguation was excluded from analysis since only two empirical paper (specifically, Milojević, 2013; Yoshikane et al., 2009) has used it for author name disambiguation.

In algorithmically disambiguated networks KISTI, MAG, and MEDLINE, a unique author identity is represented by a unique alpha-numeric descriptor assigned by researchers who performed the disambiguation in each data. In DBLP, a unique author is distinguished by an alphabetical name string, sometimes followed by a four-digit number to distinguish homonyms. In coauthorship networks disambiguated by name initials, a unique author identity (vertex) is represented in the text string format of a full surname, a comma, a shift and given name initial(s).

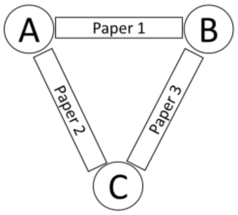
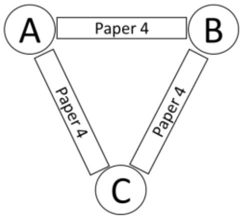
In KISTI and MEDLINE, surname and given name tokens are clearly distinguished by the data providers. In DBLP and MAG, however, name strings are recorded in the order of given name(s) and a surname without any delimiter. To apply initial based disambiguation to these data, the surname part of each name string needs to be identified. This thesis follows the method described in Kim, Kim, et al. (2014) where surnames were automatically detected using the rules of surname decisions learned from a sample of about 400,000 names recorded in journal papers in the domain of computer science as indexed in Web of Science. The accuracy of surname detection was tested through 10 samples of 100 names by checking the CVs, personal blogs, or institutional webpages associated with the scholar name. The rules' performance was accurate by on average 96.3%.

4.2 ONE-MODE VS TWO-MODE NETWORKS

In each coauthorship network, two author identities (whether they are distinguished by algorithmic, first-initial, or all-initials name disambiguation) that appear in a paper's byline were connected by edges. Here, network vertices are of the same type: in other words, vertices represent author identities. This is a one-mode network. Following prior studies, only the existence of edges was considered, while their frequency was ignored. This produced undirected, binary networks. Most coauthorship network measures such as degree centrality or average shortest paths have been applied to one-mode, binary networks.

Recently, several scholars have begun to question whether the one-mode network approach can properly represent coauthorship networks. In particular, such concerns have been raised about clustering coefficient. According to Newman (2001a) and Opsahl (2013), a coauthorship network is originally a two-mode network, where papers and authors constitute two different types of vertices and connections only exist between papers and authors, not between papers or between authors. Such a two-mode network can be converted onto (1) a paper-by-paper network, where two papers are linked by a co-sharing relationship if those papers are connected to the same authors, or (2) an author-by-author network, where two authors are linked by a coauthoring relationship if those authors are connected to the same paper(s). The one-mode coauthorship network corresponds to the latter. A problem is that this projection creates artefactual clustering of vertices. For example, as shown in Table 4, clustering of vertices is supposed to refer to Case 1: vertex A is connected to both vertex B and vertex C. Then, vertex B and vertex C that share vertex A are linked by an edge. This process is called triadic closure (Opsahl, 2013), a.k.a. transitivity (Newman, 2001b). As shown in Case 2, a one-mode network projected from a two-mode network with three authors being connected to a single paper can create the same triadic closure as in Case 1, although a real triadic closure does not happen. This bias can lead to finding inflated rates of clustering of vertices in coauthorship networks.

Table 4: Illustration of Artifact Clustering (A, B, and C represent authors; adopted from Kim and Diesner (2017))

Case No.	Input Data	Visualization of Projected One-mode Network
Case 1	Paper 1: A and B Paper 2: A and C Paper 3: B and C	
Case 2	Paper 4: A, B, and C	

To correct such errors, scholars have suggested that clustering of vertices should be measured on a two-mode network (Newman, Strogatz, & Watts, 2001; Opsahl, 2013). In this thesis, the clustering coefficient for triadic closure is measured on its two-mode network as well as its one-mode network. For this purpose, an author is linked to a paper if she appears in the author list (i.e., byline) of the paper. The list of such author-paper pairs will constitute edge lists for a two-mode network.

4.3 RANDOM NETWORKS

Scholars have compared properties of empirical networks against random networks having the same or similar properties as the target empirical networks to obtain a deeper understanding of network topology and edge generation mechanism (e.g., Perc, 2010; Robins, Pattison, Kalish, & Lusher, 2007; Watts, 1999). One of this thesis' aims is to investigate whether author name disambiguation errors affect our understanding of emergence of local network patterns (e.g., triangles) or network topologies (e.g., scale-free). For this purpose, we should know first whether each network obtained from datasets disambiguated by algorithms or name initials rather follows properties different from random networks. In other words, if two networks from the same dataset but disambiguated by algorithms and given name initials, respectively, are found not to coincide with networks generated by a random process and shown to result in different properties and topologies from random networks, we can say that name disambiguation may affect network properties. For this task, classical (a.k.a. Bernoulli or Erdős–Rényi) random networks were generated for comparing difference in statistical properties, local patterns, and topologies of computationally disambiguated and ambiguous networks.

4.3.1 One-Mode Random Networks: These networks were generated by first finding the number of vertices in a target empirical dataset. Each pair of the vertices was, then, assigned a uniform probability of forming an edge based on the number of edges in the empirical target data. Practically, the probability is the density of the target network. An R package, *i*graph, was used to create one-mode random networks that have the same or similar number of vertices and edges as empirical networks studied in this thesis.

4.3.2 Two-Mode Random Networks: The same procedure as for creating one-mode random networks can be applied for generating two-mode random networks. First, the numbers of primary (authors) and secondary (papers) vertices are found from an empirical (two-mode) target network. Then, a uniform probability of forming an edge is assigned to pairs of primary and secondary vertices. Unlike the one-mode random networks, the probability here uses the number of edges divided by the product of the numbers of primary and secondary vertices. Next, the two-mode random networks are projected onto one-mode networks for calculating common network measures. An exception is that the clustering coefficients (i.e., transitivity) for two-mode networks are calculated on the random networks without projection. This procedure was implemented by using an R package, *t*net.

4.4 MEASUREMENT

The impact of name disambiguation was measured by calculating various network metrics as follows. Measures were selected for the purpose of comparison because previous coauthorship network studies have used them. If a network metric could be calculated in two or more ways, the approach widely used in previous studies was selected. Network metrics were calculated mainly by the R package *i*graph (Csardi & Nepusz, 2006). For the calculation of clustering in two-mode networks, the R package *tnet* (Opsahl, 2009) was used. For power-law fitting of degree distribution, the Python package *powerlaw* (Alstott, Bullmore, & Plenz, 2014) and Pajek (De Nooy, Mrvar, & Batagelj, 2011) were chosen.

4.4.1 Misidentification Rate (M-Rate): "This calculates how many unique identities in the proxy of ground-truth data have been misidentified by IBD (initial-based disambiguation – inserted in this thesis). A unique author is misidentified if his/her identity is merged with other identities and/or is split into two or more identities…The misidentification rate of an initial-based method is the ratio of author name clusters in the proxy of ground-truth data that contain an author name belonging to other identities or that fail to contain an author name that should belong to the cluster *over* the total of author name clusters (Milojević, 2013). If expressed in terms of a 'cluster F1' metric, the M-rate corresponds to (1—Cluster Recall)" (Kim & Diesner, 2016).

4.4.2 Number of Vertices: This is the number of unique author identities. This corresponds to the number of author name clusters that are distinguished by algorithmic, first-initial, and all-initials name disambiguation methods.

4.4.3 Production: An author's production is the number of papers produced by the author. In practice, this corresponds to summing up the frequency of an author's unique IDs or names in data. The average production is reported for a network.

4.4.4 Gini Coefficient: The inequality in production among unique author identities is measured by the Gini coefficient, as in other coauthorship network studies (e.g., Franceschet, 2011; Martin et al., 2013; Yoshikane & Kageura, 2004). In this thesis, I chose the method by Glasser (1962) to calculate the Gini coefficient as follows:

$$G = \frac{1}{2\mu n^2}\sum_{i=1}^{n}(2i \ - n - 1)X_i \qquad (n > 1)$$

"Here, X_i is the publication frequency of an author identity X sorted from smallest to largest, *n* is the total publication frequency of all author identities observed, and μ is the mean frequency. The value of *G* can range from 0 (all authors have the same number of papers) and 1 (one author published all of the papers)" (Kim et al., 2016).

4.4.5 Number of Unique Edges: An edge in a coauthorship network represents the existence of collaboration relationship between two authors. Self-loops and multiple edges between two vertices are ignored. Only the existence of a connection between pairs of vertices is considered as unique edges for analysis.

4.4.6 Degree: Degree (or degree centrality) counts the number of vertices neighboring a target vertex. In a coauthorship network, an author's degree means the number of unique coauthors of the author. In a network matrix, degree centrality ($\mathrm{A_D}$), is denoted as an equation below (Diesner, Evans, & Kim, 2015). Here, $\mathrm{x_{ij}}$ expresses an edge between vertex i and j and ignores its frequency.

$$\mathrm{A_D(i)} = \sum_{\mathrm{j=1}}^{\mathrm{N}} \mathrm{x_{ij}(i \neq j)}$$

4.4.7 Degree Distribution: Scholars have often used coauthorship networks to test degree distribution of vertices to see if a power-law can characterize the topology of a network (e.g, Barabási et al., 2002; Milojević, 2010; Newman, 2001b). The power-law distribution of vertex degree in networks is the probability distribution of vertices (authors) with an *x* degree as follows in a simplified form,

$$p(x) = x^{-\alpha}$$

Once the power-law slope (α) on a log-log plot of degree distribution has been found, plausible mechanisms that generate such a distribution are proposed. According to preferential attachment which is one of such proposed mechanisms, for example, vertices in a network have a tendency to aim to connect to vertices with high degree centrality and this effects grow over time, generating a power-law distribution of vertex degree (Barabási et al., 2002; Milojević, 2010). Although criticisms of the measurement and utility of power-law fitting exist (e.g., Clauset, Shalizi, & Newman, 2009; Stumpf & Porter, 2012), power-law distribution of degree centrality in scientific collaboration has been cited as one of the impact findings in bibliometrics (Barabási & Frangos, 2014; Newman, 2010).

Following the suspicion of Fegley and Torvik (2013) and Kim and Diesner (2015) that power-law distribution in coauthorship networks may be an artifact due to author name disambiguation errors, this thesis attempts to fit degree distributions of networks to the power-law distribution. For this purpose, a fitting algorithm using the "maximum-likelihood fitting method with a goodness-of-fit test based on the Kolmogorov-Smirnov statistic" (Clauset et al., 2009) was used.

Another approach to testing if a network's degree distribution follows a power law is to generate a synthetic network that has the same network properties as the empirical target network and follows a power-law distribution of degree. This approach allows us to compare two distributions (one from the empirical network and the other from the synthetic network) for any discrepancy in plots (Kim & Diesner, 2015; Kim et al., 2016). The power-law-abiding synthetic networks were generated by *Pajek* (De Nooy et al., 2011).

4.4.8 Density: Network density is calculated as the ratio of the number of existing edges over the number of potential edges among all vertices in a network.

4.4.9 Centralization: This measures how degree centrality values in networks are concentrated or varied (A_D) (Wasserman & Faust, 1994), defined as:

$$C_D = \frac{\sum_{i=1}^{N}(\max(D) - D_i)}{\max \sum_{i=1}^{N}(\max(D) - D_i)}$$

The denominator refers to "the theoretically maximal sum of differences (taken pairwise between vertices)" in degree centrality (Diesner, Evans, & Kim, 2015).

4.4.10 Ratio of the Largest Component: A network component is defined as a set of vertices where each vertex can connect to others via one or more steps of connections. The ratio of the number of vertices belonging to the largest component over the total of all vertices in a network is reported.

4.4.11 Average Shortest Path Lengths: The shortest path length, or the geodesic, between two vertices in a component is the minimum number of edges that links them. The average shortest path lengths of pairs of vertices that are reachable in each dataset (Brandes, 2001) were calculated.

Counting the average shortest path lengths can be time- and memory-consuming for a large-scale network as the time complexity is known to increase by $O(|V| + |E|)$, where V stands for the number of vertices and E for the number of edges (Fegley & Torvik, 2013). Thus, this thesis estimated the average shortest path lengths for networks with more than 500,000 vertices or edges. Although various methods for estimating the average shortest paths have been proposed (e.g., Illenberger & Floetteroed, 2012; Potamias, Bonchi, Castillo, & Gionis, 2009), this thesis follows the approach by Fegley and Torvik (2013), where a set of 1,000 randomly sampled vertices is used as estimation points. Specifically, the average shortest path lengths from these 1,000 vertices to all the other vertices are calculated as a proxy for the average shortest path lengths for the whole networks.

To measure the accuracy of this approach, the average shortest path lengths for the whole network generated from computationally disambiguated DBLP was obtained (= 6.49). Then, from this network, a sample of 10, 100, and 1,000 vertices were randomly chosen for 10 times each. Next, average shortest path lengths were calculated from these selected vertices to all other reachable vertices. Table 5 summarizes the results of this estimation procedure. Here, the second row in the table reports the average shortest paths with standard deviations. In the third row, the average absolute difference between the ground-truth (=6.49) and estimated values is shown with the corresponding error ratio, the average error divided by the ground-truth. The average shortest path lengths with 1,000 sampled vertices approximated the ground-truth with an average error ratio of 0.27%. In this thesis, sets of 1,000 sampled vertices were used for estimating average shortest path lengths.

Table 5: Results of Estimating Average Shortest Path Lengths

Estimation	Number of Sampled Nodes		
	10	100	1,000
Avg. Shortest Path Lengths (Standard Deviation)	6.59 (0.28)	6.52 (0.08)	6.50 (0.02)
Average Error (Average Error Ratio)	0.24 (3.72%)	0.07 (1.10%)	0.02 (0.27%)

4.4.12 Degree Assortativity: "This measures the extent to which unique authors collaborate with others who are similar to them in terms of degree centrality" (Kim & Diesner, 2016; Kim, Kim, & Diesner, 2014). Technically, degree assortativity is measured as "the Pearson correlation coefficient of the degrees at either ends of an edge" between pairs of vertices (Newman, 2002).

4.4.13 Clustering Coefficient (Transitivity): A network configuration is a subgraph that represents local patterns in a network (Robins et al., 2007). Among possible network configurations, triangles have attracted special attention from coauthorship network scholars. In the context of scientific collaboration, a triangle implies that two scholars who did not collaborate previously but worked with a shared scholar tend to collaborate with each other later. This may happen because the shared scholar introduces those two scholars to each other. This tendency in a coauthorship network has been measured by the global clustering coefficient, a.k.a. transitivity (Newman, 2001b).

Technically, this transitivity (NCC) refers to ratio of triadic closure among author identities: “the probability of forming an edge between two vertices that have a common neighbor” (Newman, 2001b), which is expressed as follows (Fegley & Torvik, 2013):

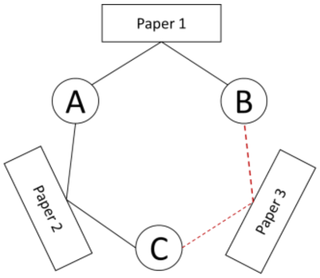
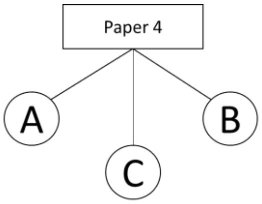
$$NCC = 3 \times \frac{\text{number of triangles on the network}}{\text{number of connected triples of vertices}}$$

As discussed above, clustering of three vertices can be inflated due to the projection of a two-mode network onto a one-mode network. In order to correct such errors, Opsahl (2013) proposed a clustering coefficient (OCC) in a two-mode network expressed as follows:

$$OCC = \frac{\text{number of closed 4paths}}{\text{number of 4paths}}$$

In Table 6, the 4path in Case 1 refers to the sequence of edges ‘Vertex C - Paper 2 – Vertex A - Paper 1 – Vertex B.’ This 4path is closed by the edge ‘Vertex B – Paper 3 – Vertex C.’ In contrast, the three vertices in Case 2 do not show a triadic closure according to this measure. Since most previous studies only report Newman’s measure of clustering, this thesis reports both a) Newman’s measure to be comparable to previous studies and b) Opsahl’s measure to provide a more correct representation of clustering. The transitivity (clustering coefficient) of a one-mode network was measured by using *i*graph, while that of a two-mode network was measured by using *t*net.

Table 6: Illustration of Opsahl (2013)'s Measure of Clustering (reprinted from Kim & Diesner (2017))

Cases	Input Data	Visualization of Clustering of Three Vertices
Case 1	Paper 1: A and B Paper 2: A and C Paper 3: B and C	
Case 2	Paper 4: A, B, and C	

4.4.14 k-2-paths: A network configuration can be used to infer edge formation mechanism in a network. Specifically, if a particular network configuration appears in a network more frequently than expected by chance, the configuration can be said to have a propensity to be prevalent in the network (Shumate & Palazzolo, 2010). This tendency can be tested by comparing the frequency of a target configuration in an empirical network and in random networks simulated based on the empirical network that has the same or similar number of vertices and edges.

In this thesis, the prevalence of *k*-2-paths was tested. As shown in Figure 1, 2-paths refer to a dyad of vertices that are not linked to each other but to third vertices from one to *k*. This configuration has been often studied by link prediction scholars (e.g., Guns & Rousseau, 2014; Liben-Nowell & Kleinberg, 2007; E. Yan & Guns, 2014). Specifically, the larger the number of shared coauthors between two authors is, the higher the probability they form a collaboration edge (Kim & Diesner, 2017; Newman, 2001a). This hypothesis has been tested by calculating ratios of the closed 2-paths (e.g., an edge formed between X and Y in the figure) over the open 2-paths (e.g., non-edge between X and Y in the figure) as a function of the number of shared coauthors (1…*k*).

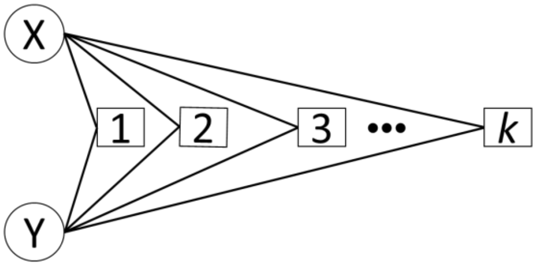


Figure 1: Illustration of k-2-Paths

The prevalence of *k*-2-paths, where a dyad is not linked despite the large number of shared *k* vertices, can lead to the finding that authors in a coauthorship network tend not to collaborate with each other even if they have many coauthors in common. To find out how name disambiguation methods affect the prevalence of *k*-2-paths in the same networks, the numbers of *k*-2-paths were counted for *k* = 1…15 by using an R package, *statnet*. Due to the consistent degeneracy of statistical modeling of the configuration, two-mode random networks were generated per disambiguation method in each dataset and counts of *k*-2-paths were obtained for comparing them to the empirical networks where author names were distinguished by algorithmic and initial-based methods.

CHAPTER 5: RESULTS

5.1 OVERVIEW OF IMPACT OF MERGING AND SPLITTING

In this chapter, the impact of merging and splitting will be discussed in detail[7]. To help readers understand better, a simple case is illustrated by Figure 2 and Table 7.

Let's assume that there are two papers coauthored by three unique authors each (left figure). During the initial name disambiguation phase, two authors are merged into one identity ("Kim, June" and "Kim, Jay" into "Kim, J.") as shown in the right figure below. If two author identities are merged, the number of unique authors decreases. The number of edges does not change (see the 4th row of the Table 7), but other network metrics change significantly. Depending on the measure, the direction of change is positive (increase) or negative (decrease). The impact of splitting is a reversal of the merging effect (i.e., changes from right sub-figure to left sub-figure).

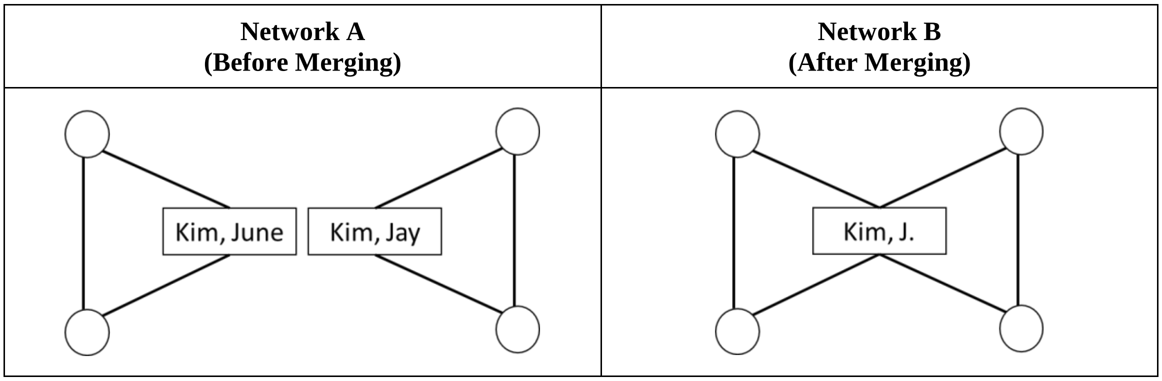


Figure 2: An Example of Simple Merging Scenario

[7] I rely heavily in this chapter on Kim and Diesner (2015), Kim and Diesner (2016), Kim, Kim, and Diesner (2014), and Kim, Diesner, Kim, Aleyasen, and Kim (2014).

Table 7: Summary of Network Property Change for an Illustrative Case in Figure 2

Metrics	Network A	Network B	Change (%)
Number of Vertices	6	5	-17%
Avg. Productivity	1.00	1.20	+20%
Number of Edges	6	6	0%
Density	0.40	0.60	+50%
Avg. Degree	2.00	2.40	+20%
Largest Component Size	3 (50%)	6 (100%)	+100%
Transitivity	1.00	0.60	-40%
Assortativity	N/A	-0.5	-
Avg. Shortest Paths (only calculated for reachable vertices)	1.00	1.40	+40%

5.2 MISIDENTIFICATION RATE

As a preliminary step to understanding the impact of name disambiguation on findings in bibliometric data, the amount of errors in author identification by initial-based disambiguation needs to be calculated. This is important because the types and levels of misidentification will affect the interpretation of findings throughout this study. Figure 3 shows the change in the misidentification rates over time. A misidentification rate is the number of merged and/or split authors by first-initial or all-initials method over the total numbers of unique authors identified by algorithmic disambiguation. Specifically, an author identity in algorithmically disambiguated data is assigned to one of four categories depending on the compromising type: (1) Type A: no merging and/or splitting (Blue), (2) Type B: merging only (Orange), (3) Type C: splitting only

(Black), and (4) Type D: both merging and splitting (Yellow). In each subfigure, the ratio of each type (value is set between zero and one) is depicted on y-axis over years (x-axis).

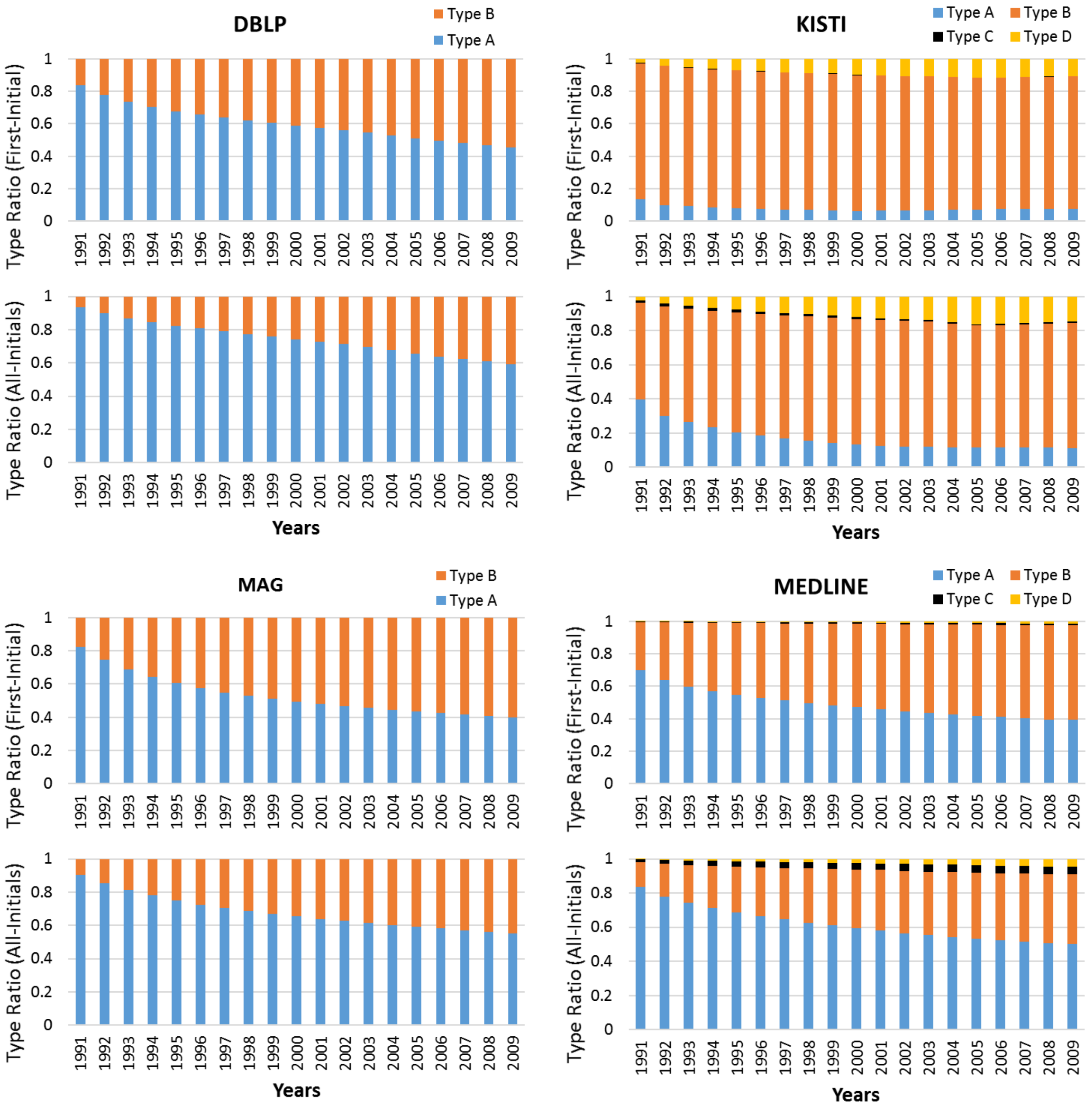


Figure 3: Trend of Misidentified Authors by Initial-Based Disambiguation

First, the ratio of unique author identities merged or split by initial-based methods is not negligible. For example, in DBLP, 54% of unique author identities in algorithmically disambiguated data have been merged by the first-initial method, and about 41% of them get merged by the all-initials method (for 2009). In KISTI, the misidentification rates increase up to 82% by the first-initial method and 73% by the all-initials method for 2009. In KISTI, this phenomenon is extreme because more than 97% of the names in KISTI are Korean, where people often share the same surnames and given names. Overall, the misidentification rates keep increasing over time in all datasets. This can be because, as new names are added to data, some of them are likely to match previously uncompromised names in the data in terms of surname and initialized given names, which leads to merging or splitting.

As one might expect, in all datasets over time, all-initials method performed consistently more accurately than the first-initial method. In other words, the ratio of compromised author identities by the all-initials method is lower than that by the first-initial method. This observation is contrary to the finding by Milojević (2013) where the first-initial disambiguation was “superior” to the all-initials disambiguation in detecting “true” author identities. However, the observation is in line with Newman (2001b), who argued that the all-initials name disambiguation method provides the ceiling of the number of unique author identities in scholarly data, while first-initial method produces the bottom limit. The finding is expected as the all-initials method adds more detail to name strings for identity match and, thus, helps name ambiguity resolution.

Regarding types, Type B (merging only) is dominant with both the first-initial and the all-initials methods in all datasets over all years. In other words, Type B happened far more often than Type C (splitting only) and Type D (both merging and splitting). Another noticeable observation for types is that, while Type C and Type D occur in KISTI and MEDLINE, they do not occur in

DBLP and MAG. This does not mean that Type C and Type D did not happen in DBLP and MAG. Instead, this is related to the characteristics of name disambiguation in these datasets. First, a unique author in DBLP is represented by an alphabetical name string (sometime followed by a four-digit identification number). This means that in DBLP, a unique author is not allowed to have two or more name strings. An instance of splitting happens when an author has two or more name variants that are different when given names are initialized. Therefore, splitting by initial-based disambiguation cannot occur in DBLP where an author is assigned a unique name string for her/his identification.

Second, a unique author in MAG is represented by a unique alpha-numeric string. As noted in 3.3 MAG in Chapter 3, algorithmic disambiguation in MAG was not perfect and produced compromised author identities. Especially, splitting was pronounced when MAG's disambiguation performance was tested against ground-truth data (see Table 3). This means that many unique author IDs in MAG were algorithmically assigned without correcting splitting error. In addition, MAG has no Type C and Type D when disambiguated by initial-based method. This implies that, during the algorithmic disambiguation for MAG, two names that are different in strings were not compared for possible identity matching. This corroborates the conjecture above that MAG did not deal with splitting properly and shows that the disambiguation algorithm for MAG was not as sophisticated as that for MEDLINE and KISTI which implemented a splitting-correction procedures.

5.3 NUMBER OF UNIQUE AUTHORS

One of the basic questions we can ask about a publication dataset is how many unique authors it records and how this number changes over time. The answer to this question matters because it helps us to estimate the size and growth of scientific communities recorded in bibliometric data and serves as the basis for calculating other metrics such as average production per author. Figure 4 shows the temporal change in the number of unique authors for three name disambiguation methods: algorithmically disambiguated (circles), first-initial based (triangles), and all-initials based (crosses) methods. The number of unique authors is marked up to the target year for each disambiguation method. An inset figure shows the ratio of error calculated as follows.

$$Error\ Ratio\ (\%) = \frac{Val_A - Val_B}{Val_B} \times 100$$

Here, Val_A refers to the number of unique authors distinguished by first-initial or all-initials disambiguation, while Val_B refers to the number of unique authors detected by algorithmic disambiguation.

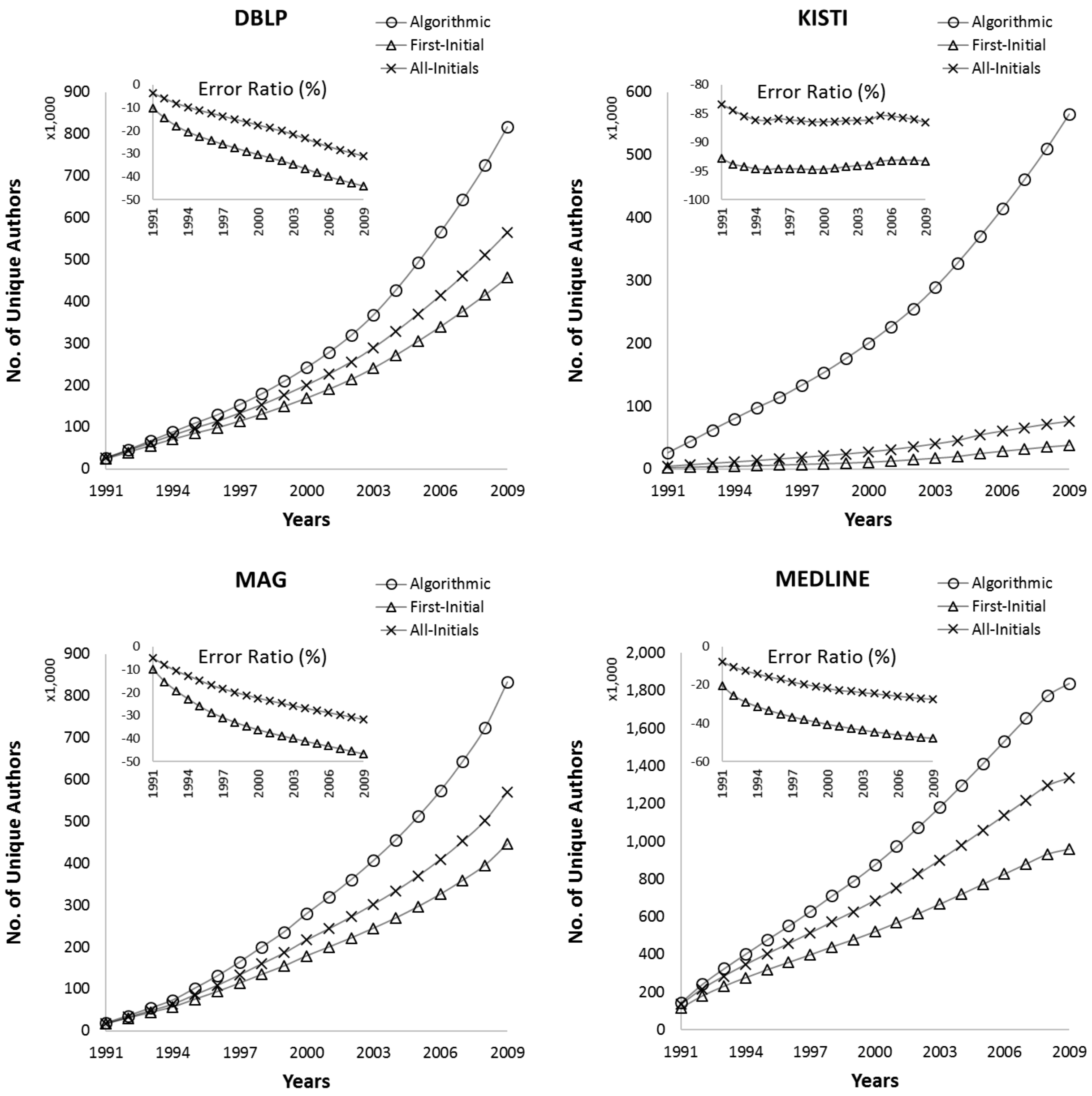


Figure 4: Trend of Number of Unique Authors

Overall, the trendlines for all four datasets show an increase in the number of unique authors over years. Specifically, the trendlines for the number of unique authors follow either an exponential or a linear growth over time for algorithmic, first-initial, and all-initials author name disambiguation methods in all datasets. Table 8 summarizes the results of fitting an exponential curve or a linear line to each dataset and reports the best fitting model, equation, and R-squared value for fit. If converted into a log (for *y*-axis)-linear (for *x*-axis) graph, for example, the growth plot of the number of unique authors by the algorithmic method in DBLP can be fitted with an exponential function (exponent = 0.17; R-squared = 0.96), while other plots of unique authors by initial-based disambiguation better fit a linear line. Except DBLP, the results in the table indicate that all three disambiguation methods suggest the same type of growth pattern for KISTI (exponential), MAG (linear), and MEDLINE (linear).

Table 8: Trendline Fitting for Number of Unique Authors

Data	**Disambiguation Method**		
	Algorithmic	**First-Initial**	**All-Initials**
DBLP	Exponential $y = 6E{-}142e^{0.17x}$ $R^2 = 0.96$	Linear $y = 23104_x - 5E07$ $R^2 = 0.96$	Linear $y = 28644_x - 6E07$ $R^2 = 0.96$
KISTI	Exponential $y = 1E{-}126e^{0.15x}$ $R^2 = 0.95$	Exponential $y = 3E{-}135e^{0.16x}$ $R^2 = 0.99$	Exponential $y = 1E{-}123e^{0.15x}$ $R^2 = 0.97$
MAG	Linear $y = 42895_x - 9E07$ $R^2 = 0.96$	Linear $y = 22863_x - 5E07$ $R^2 = 0.98$	Linear $y = 29376_x - 6E07$ $R^2 = 0.97$
MEDLINE	Linear $y = 94359_x - 2E08$ $R^2 = 0.99$	Linear $y = 46331_x - 9E07$ $R^2 = 1.00$	Linear $y = 66546_x - 1E08$ $R^2 = 1.00$

A noticeable observation is that initial-based name disambiguation underestimates or deflates the number of unique authors for all years in all datasets. In other words, the plots for algorithmic disambiguation (circles) appears consistently above those for first-initial (triangles) and all-initials (crosses) disambiguation. This implies that “merging of author identities happens more often than splitting (merging reduces the number of unique identities while splitting increases it)” by initial-based disambiguation (Kim & Diesner, 2015). This is consistent with the observation from Figure 3 that merging happened more frequently than splitting (for KSITI and MEDLINE) or that only merging happened (DBLP and MAG). The merging effect is most pronounced in KISTI: initial-based methods underestimate or deflate the numbers of unique authors by 85%~95% (see the inset figure – Error Ratio – for KISTI). This is because the majority of Korean people share a small set of surnames and given names, which increases the ambiguity to a level where only a small portion of names remain unique after initial-based disambiguation.

The underestimation by initial-based disambiguation leads us to challenges the long-held assumption in coauthorship network studies 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" (Kim & Diesner, 2015). This assumption was first proposed by Newman (2001b). Since then, scholars analyzing scholarly data have based their argument on this assumption that they can estimate properties of a correctly disambiguated network to exist between properties of two networks pre-processed by the first-initial and all-initials name disambiguation methods, respectively (e.g., Barabási et al., 2002; Milojević, 2013; Newman, 2001b; Wagner & Leydesdorff, 2005). Figure 4 shows that the number of unique authors detected by algorithmic disambiguation is found beyond the upper bounds (the upper bounds represent the largest numbers of unique authors distinguished by the all-initials method). Recent studies (Fegley & Torvik, 2013; Kim & Diesner, 2015, 2016) also found this "off-upper-bound phenomenon" in a static analysis of network data from MEDLINE (2003-2007), USPTO (2003-2007), and Web of Science (2012) as well as a temporal analysis of DBLP (1984-2013) data.

Another observation is that the gaps between trendlines by initial-based method and algorithmic method have kept increasing over time. For example, in 1991, algorithmic disambiguation identified 146K authors in MEDLINE, while the first-initial method found 116K (-20% compared to algorithmic disambiguation) and all-initials method found 135K (-8% compared to algorithmic disambiguation). In 2009, these numbers are 1.8M by algorithmic disambiguation, 960K by first-initial (−48%), and 1.3M by all-initials (−27%). This trend can be confirmed by the yearly trend of error ratio in the inset figures, where the size of error in all datasets increases over the years. If this trend continues, “the prediction of the size of a scientific community…can be very different” depending on the name disambiguation method (Kim & Diesner, 2015).

5.4 AVERAGE PRODUCTION

How is the identification error affecting our understanding of author production in scholarly data? Figure 5 shows the temporal change of average author production in four datasets per disambiguation method: algorithmic disambiguation (circles), first-initial method (triangles), and all-initials based one (crosses). The average production is marked up to the target year for each disambiguation method. An inset figure shows the ratio of error calculated as follows.

$$Error\ Ratio\ (\%) = \frac{Val_A - Val_B}{Val_B} \times 100$$

Here, Val_A refers to the average production calculated for authors disambiguated by first-initial or all-initials based disambiguation, while Val_B refers to the average production of authors identified by algorithmic disambiguation.

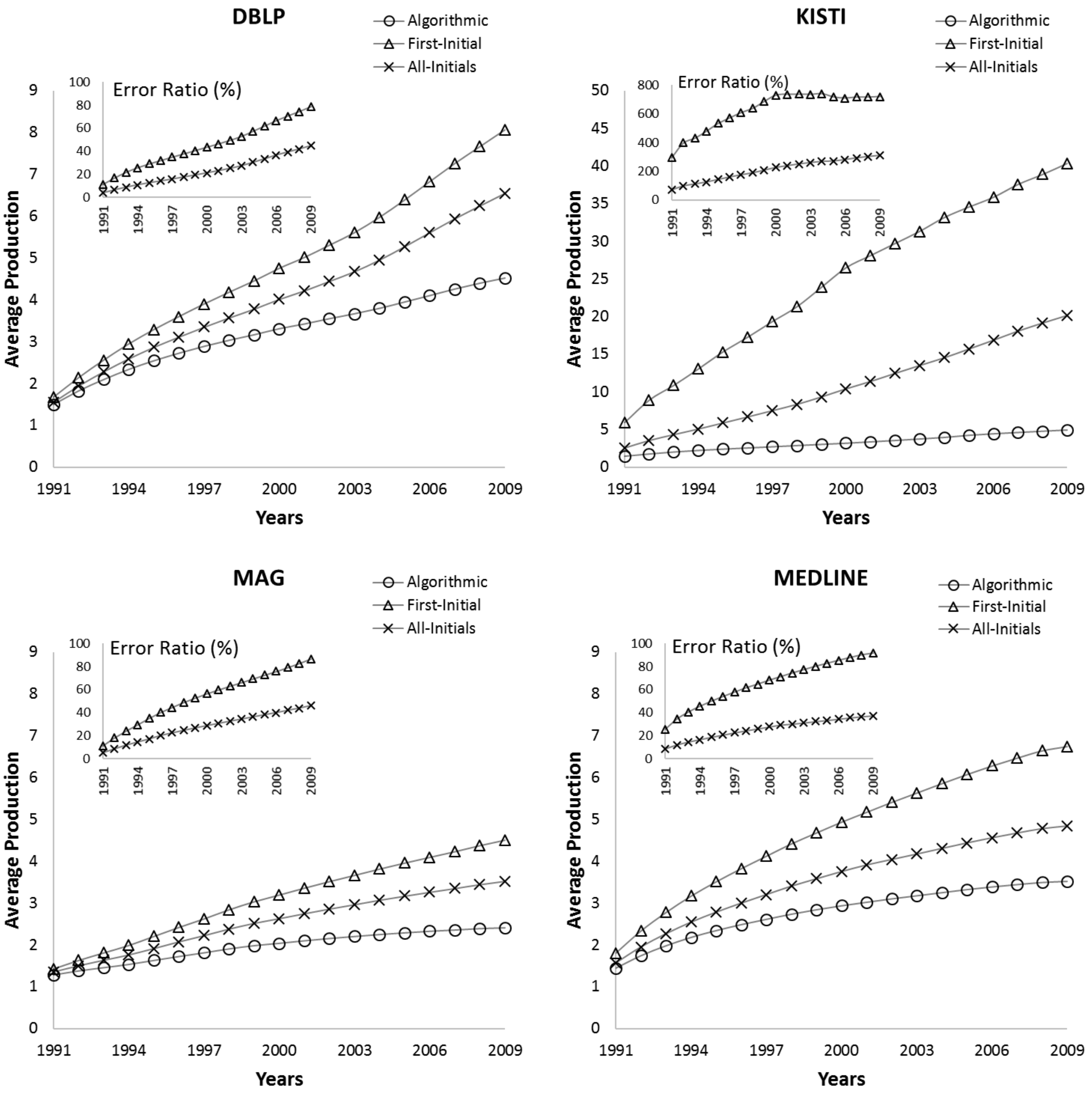


Figure 5: Trend of Average Production

Some observations are worth noting. First, average author production lines all increase over time regardless of disambiguation method. For most plots, any meaningful growth curves, e.g., exponential or linear, were not found. Second, initial-based disambiguation consistently overestimates or inflates the average production. For example, authors in algorithmically disambiguated DBLP data published on average 4.5 papers, while authors in the same data that were disambiguated by first-initial and all-initials method would produce on average 8.07 (79%↑) and 6.55 (45% ↑) papers, respectively. This is not unexpected. As multiple author identities are merged into one author, publication counts of those compromised authors are attributed to that author who becomes to have many publications. Third, the inset figures show that the error levels also increase over time, meaning the increase of error levels in estimating the number of unique authors (see insets in Figure 4). The more author identities that are merged in a single author, the more publications are assigned to them. Third, the overestimation error by first-initial method is larger than that by all-initials method as the first-initial method produces more merged identities than the all-initials method.

5.5 CONCENTRATION OF AUTHOR PRODUCTION

Several studies used the Gini coefficient to measure the inequality of the distribution of author production (e.g., Franceschet, 2011; Martin et al., 2013) Figure 6 shows the over-time change of the Gini coefficient measuring the distribution of author production in four datasets per disambiguation method: algorithmic disambiguation (circles), first-initial method (triangles), and all-initials based one (crosses). The Gini coefficient is marked up to the target year for each disambiguation method. An inset figure shows the ratio of error calculated as follows.

$$Error\ Ratio\ (\%) = \frac{Val_A - Val_B}{Val_B} \times 100$$

Here, Val_A refers to the Gini coefficient calculated for distribution found in data disambiguated by first-initial or all-initials method, while Val_B refers to the Gini coefficient of distribution identified by algorithmic disambiguation.

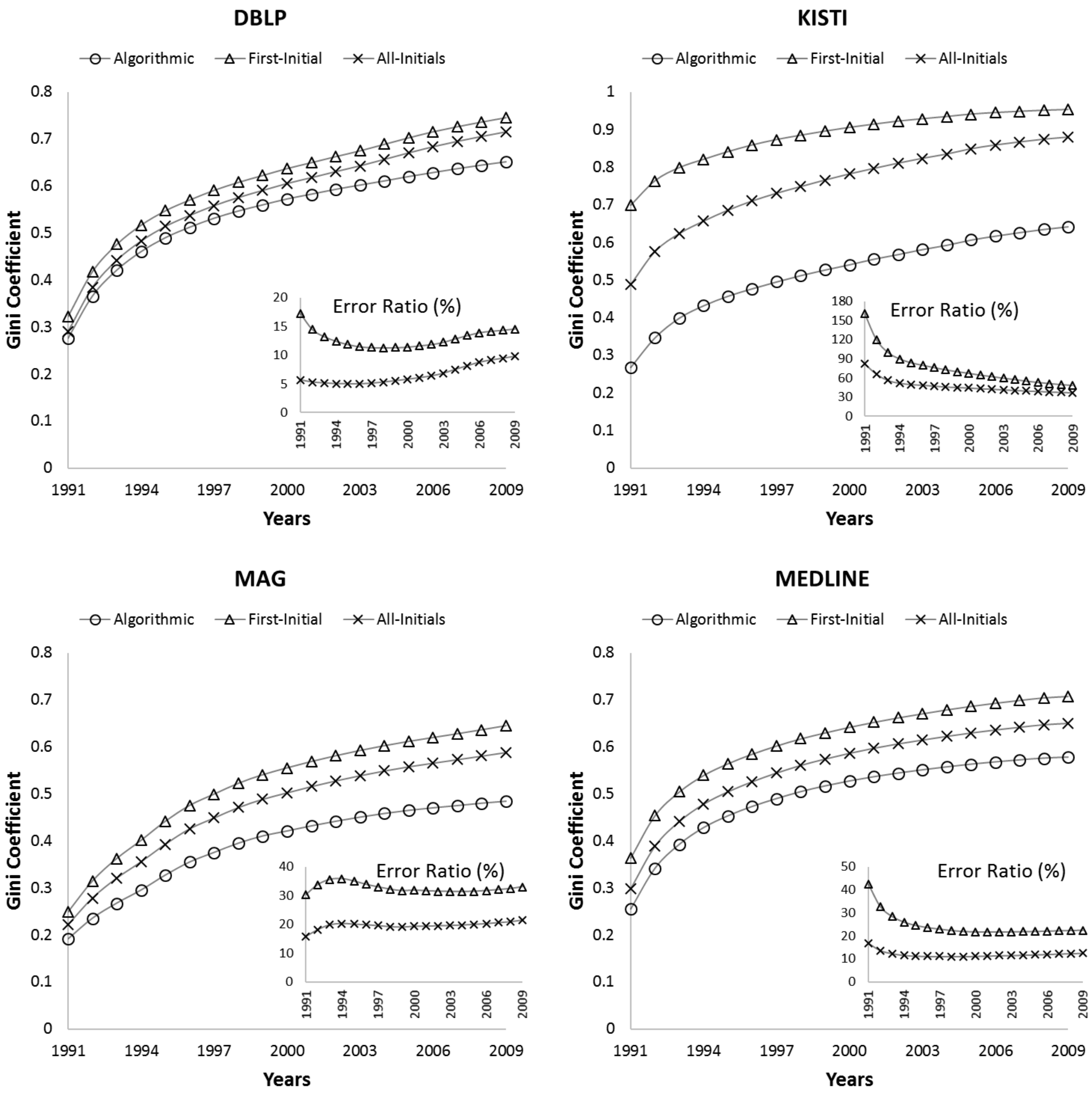


Figure 6: Trend of Gini coefficient for Author Production

“Plotting the temporal change in the Gini coefficient of” author production “reveals that inequality has been increasing over time" regardless of disambiguation method (Kim et al., 2016). This means that each method provided the same finding: some authors have managed to produce more publications than others over time. Such a tendency of unequal distribution in scientific research production has been explained by, for example, the Matthew's Effect (Merton, 1968), which states that some scholars attract more opportunities and resources for publication than others, which over time strengthens such disparity.

The initial-based method inflates the level of inequality compared to that by algorithmic disambiguation. Inset figures for Error Ratio shows that Error Ratios range from about 7% by the all-initials method in DBLP to about 75% by the first-initial method in KISTI. This difference in error level can lead to different understanding of how serious the inequality is per dataset, resulting in different policy implications, if sought.

5.6 NUMBER OF UNIQUE EDGES

In coauthorship network analysis, a unique edge represents the coauthoring activity of two scholars: two authors (nodes) are represented to be connected via an edge if they work together on a paper. Figure 7 shows the temporal change in the number of unique edges in four datasets per disambiguation method: algorithmic disambiguation (circles), first-initial method (triangles), and all-initials based one (crosses). The number of unique edges is marked up to the target year for each disambiguation method. An inset figure shows the ratio of error calculated as follows.

$$Error\ Ratio\ (\%) = \frac{Val_A - Val_B}{Val_B} \times 100$$

Here, Val_A refers to the number of unique edges calculated for data disambiguated by first-initial or all-initials disambiguation, while Val_B refers to the number of unique edges counted by algorithmic disambiguation.

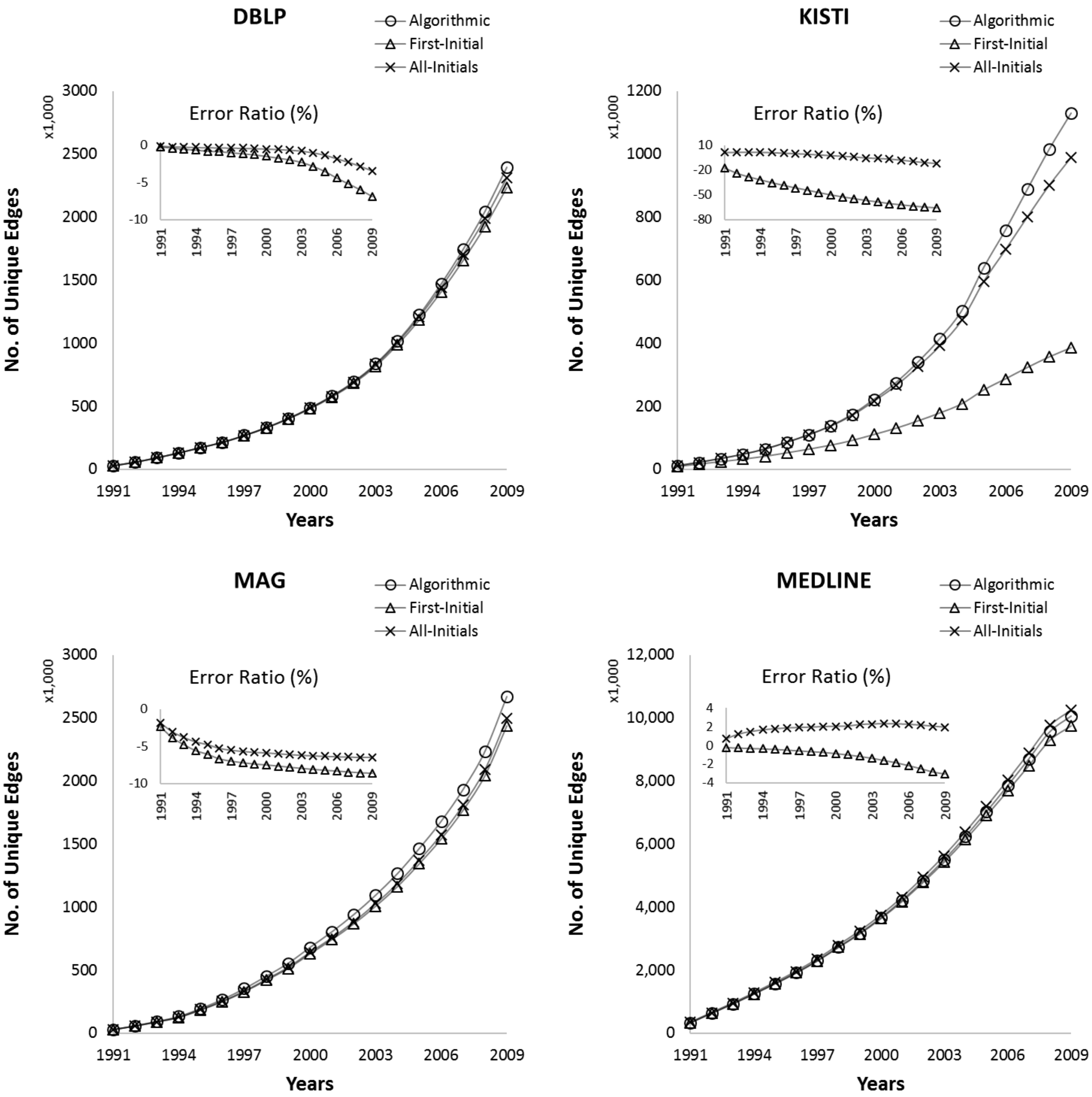


Figure 7: Trend of Number of Unique Edges

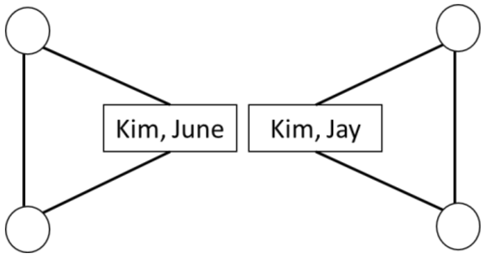
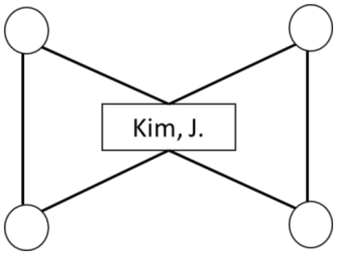
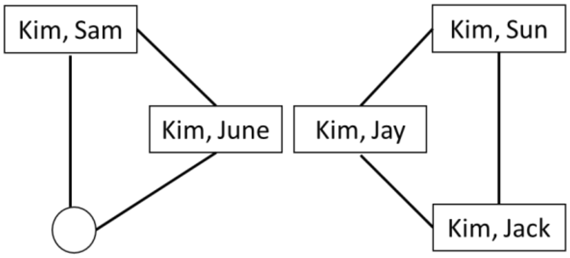
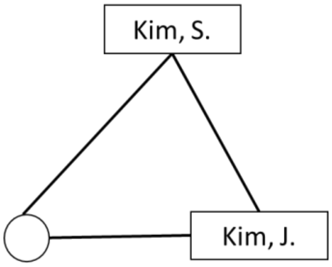
Over time, the trends of the number of unique edges demonstrate an exponential growth for DBLP, KISTI, and MAG, and a linear growth for MEDLINE. As summarized in Table 9 reporting best fitting model, equation, and R-squared fit, types of trendlines show no difference per disambiguation method in each dataset. Unlike the trends for the number of unique authors, gaps between the plots of unique edges are not pronounced except the gap between algorithmic and first-initial methods in KISTI. This can be confirmed first by the small error ratios reported in inset figures of Figure 7. In addition, in Table 9, the trendlines per dataset have similar coefficients for modeling growth.

Table 9: Trendline Fitting for Number of Unique Edges

Data	**Disambiguation Method**		
	Algorithmic	**First-Initial**	**All-Initials**
DBLP	Exponential $y = 3E{-}186e^{0.22x}$ $R^2 = 0.97$	Exponential $y = 3E{-}183e^{0.22x}$ $R^2 = 0.96$	Exponential $y = 7E{-}185e^{0.22x}$ $R^2 = 0.96$
KISTI	Exponential $y = 1E{-}205e^{0.24x}$ $R^2 = 0.97$	Exponential $y = 1E{-}163e^{0.19x}$ $R^2 = 0.96$	Exponential $y = 1E{-}198e^{0.23x}$ $R^2 = 0.97$
MAG	Exponential $y = 4E{-}190e^{0.22x}$ $R^2 = 0.94$	Exponential $y = 1E{-}187e^{0.22x}$ $R^2 = 0.94$	Exponential $y = 2E{-}188e^{0.22x}$ $R^2 = 0.94$
MEDLINE	Linear $y = 548289_x - 1E09$ $R^2 = 0.97$	Linear $y = 532991_x - 1E09$ $R^2 = 0.97$	Linear $y = 560324_x - 1E09$ $R^2 = 0.97$

This observation implies that merged authors “usually have distinct collaborators. In other words, 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, the total number of edges in the network would decrease to a noticeable extent” (Kim & Diesner, 2015). This situation is illustrated in Table 10. In Case A, two coauthorship networks are merged into one because “Kim, June” and “Kim, Jay” have the same full surname and the same first initial of given names. Although the number of unique authors decreases from six to five, the number of unique edges is not changed (i.e., six edges). In Case B, however, the first-initial method results in merging two networks into one with only three edges via two pairs of inter-network merging ("Kim, Sam" – "Kim, Sun" and "Kim, June" – "Kim, Jay") and one intra-network merging ("Kim, Jay" – "Kim, Jack"; a self-loop). The effect of splitting can be thought of as the reverse of merging: "After Merging" corresponds to "Before Splitting" and "Before Merging" to "After Splitting."

Table 10: Illustrative Cases for Merging of Edges

CASE	Before Merging	After Merging
A		
B		

The small gaps between trendlines in Figure 7 imply that the decrease by Case B in Table 10 happens at a very low level for, at least, DBLP, MAG, and MEDLINE. Thus, it can be inferred that "it is uncommon that two or more authors in a byline have ambiguous names that may lead to merging with names in other bylines for those datasets" (Kim & Diesner, 2015). In contrast, KISTI showcases that the merging of edges by Case B happens quite often. For example, in 2009 data, the number of unique edges counted by algorithmic disambiguation decreased by about

66% (first-initial method) and 12% (all-initials method). This is because many Korean authors share surnames and given names, which can cause high level of merging of author identities (as shown in Figure 4), eventually leading to frequent merging of edges (as shown by Case B in Table 10).

5.7 AVERAGE DEGREE

In network analysis, vertex degree represents the number of vertices connected to a vertex. In coauthorship networks, vertex degree represents the counts of unique coauthor identities (i.e., vertices) that has worked with an author. Figure 8 shows the temporal change of the average degree in four datasets per disambiguation method: algorithmic disambiguation (circles), first-initial method (triangles), and all-initials based one (crosses). The average degree is marked up to the target year for each disambiguation method. An inset figure shows the ratio of error calculated as follows.

$$Error\ Ratio\ (\%) = \frac{Val_A - Val_B}{Val_B} \times 100$$

Here, Val_A refers to the average degree calculated for data disambiguated by first-initial or all-initials method, while Val_B refers to the average degree obtained from data pre-processed by algorithmic disambiguation.

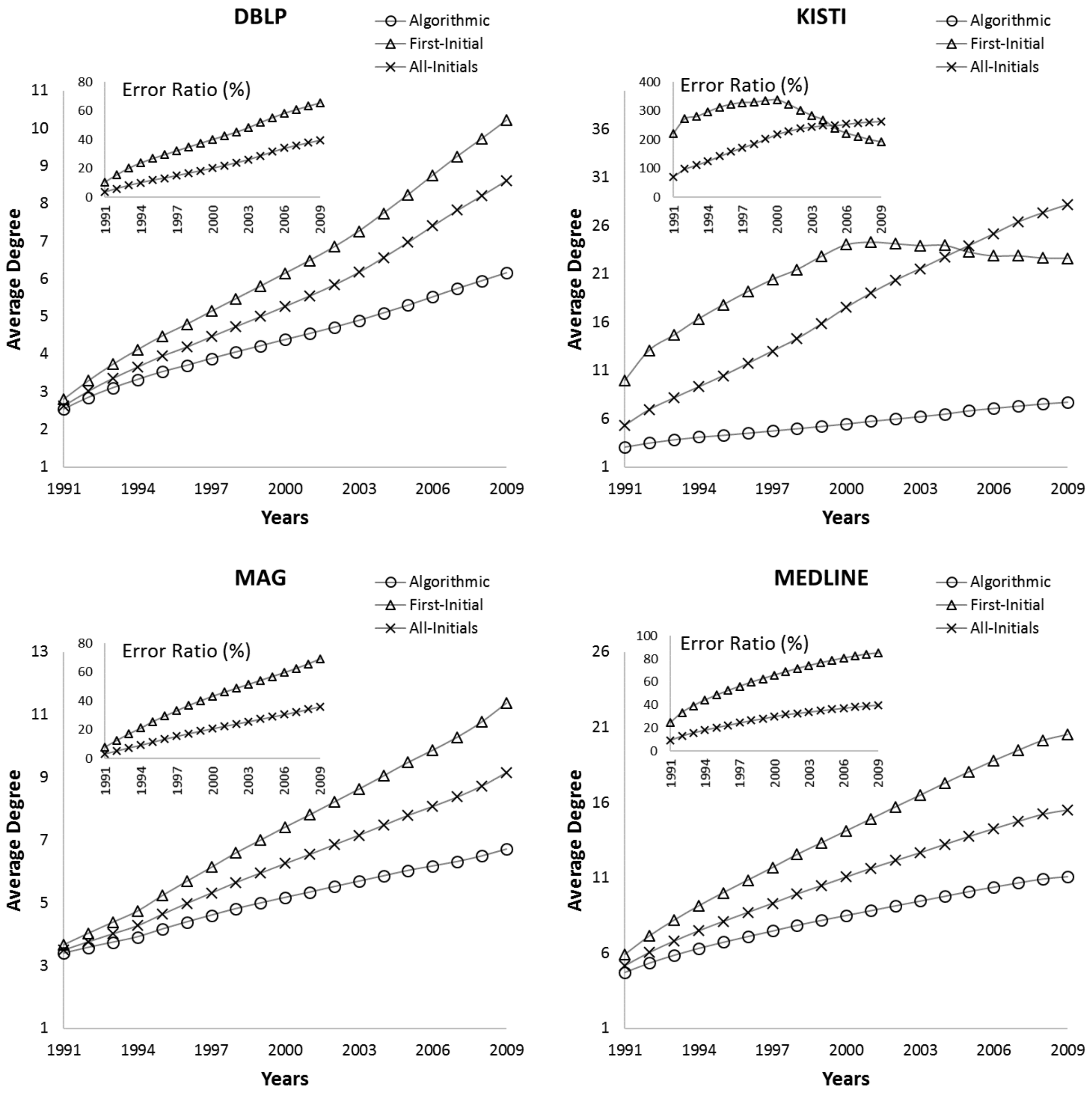


Figure 8: Trend of Average Degree

The average degree increases over time in all four datasets regardless of the disambiguation method. This tendency of scholars to work together with many coauthors has been observed for science (Waltman, 2012). An exception is the trend generated by the first-initial method in KISTI. Its average degree reached its highest point around 2000 and then continuously decreased. This may be due to the merging effect illustrated in Case B of Table 10, where both network vertices and edges are reduced in number because of name ambiguity and, accordingly, network structure is severely distorted. We can only conjecture that due to this type of merging, the structure of network disambiguated by algorithm went through a fundamental change, resulting in substantially different network properties.

The level of increase is, however, quite different per disambiguation method. Overall, initial-based disambiguation overestimated or inflates the average degrees. For example, for 2009, the average degree of authors in DBLP is 4.52 with the algorithmic disambiguation, and increases to 8.07 (79%) with the first-initial method and to 6.55 (45%) with the all-initials method. The merging of identities can explain this inflation of average degree. "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 (i.e., increase of numerator – inserted in this thesis), the number of unique authors (i.e., denominator – inserted in this thesis) decreases due to merging. These two effects erroneously inflate the average degree" (Kim & Diesner, 2015). The inset figures of Figure 8 also show that gaps between average degree plots by initial-based disambiguation and algorithmic disambiguation increase over time.

5.8 DEGREE DISTRIBUTION

Regarding degree centrality in networks, scholars have frequently investigated degree distribution to decide whether a power-law can characterize the topology of a network being analyzed (e.g., Barabási et al., 2002; Liben-Nowell & Kleinberg, 2007; Milojević, 2010; Newman, 2001b). If the distribution plot of degree in a network follows a straight line for its tail part when it is projected on a (cumulative) log-log scale pane, the network is described to have a power-law degree distribution. Then, scholars propose and test “plausible mechanisms leading to such a distribution” (Kim & Diesner, 2015). One of those mechanisms that have been often tested is preferential attachment, which refers to the tendency of scholars to select coauthors who are high in degree, i.e., who have many coauthors (Barabási et al., 2002). Such a propensity was simulated to mass over time and produce the power-law distribution of vertex degree in large-scale coauthorship networks (e.g., Barabási et al., 2002; Milojević, 2010; Perc, 2010).

To test whether name disambiguation affects our understanding of a network in terms of degree distribution, degree distributions obtained from networks disambiguated by algorithmic (blue circles), first-initial (green triangles), and all-initials (red crosses) methods are depicted on the same cumulative log-log scale for each dataset as shown in Figure 9. In each subfigure, the x-axis depicts the value of vertex degree (i.e., *x*), while the y-axis stands for the ratio (in percentage; %) of authors who have the *x* or above degree against the total of authors. The figure suggests several interesting observations. First, degree distributions are all highly skewed: a small group of authors have many coauthors, while most authors have a small number of coauthors. For example, in the algorithmically disambiguated MEDLINE (blue circles), 90% of authors have 23 or less collaborators, while the rest have from 24 up to 1,325 collaborators.

Second, “distributions obtained with first-initial method are positioned above those generated via 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” (Kim & Diesner, 2015). For example, in MEDLINE, the ratio of authors with a degree of 10 or more is 32% when disambiguated by algorithm, 42% by first-initial method, and 38% by all-initials method. This corroborates the findings from Figure 8. As author identities are merged via initial-based disambiguation, their coauthors also become the coauthors of the merged identities, increasing the average degree of authors. "This merging effect pushes the distribution plots right and upward when compared to those created using algorithmic disambiguation" (Kim & Diesner, 2015).

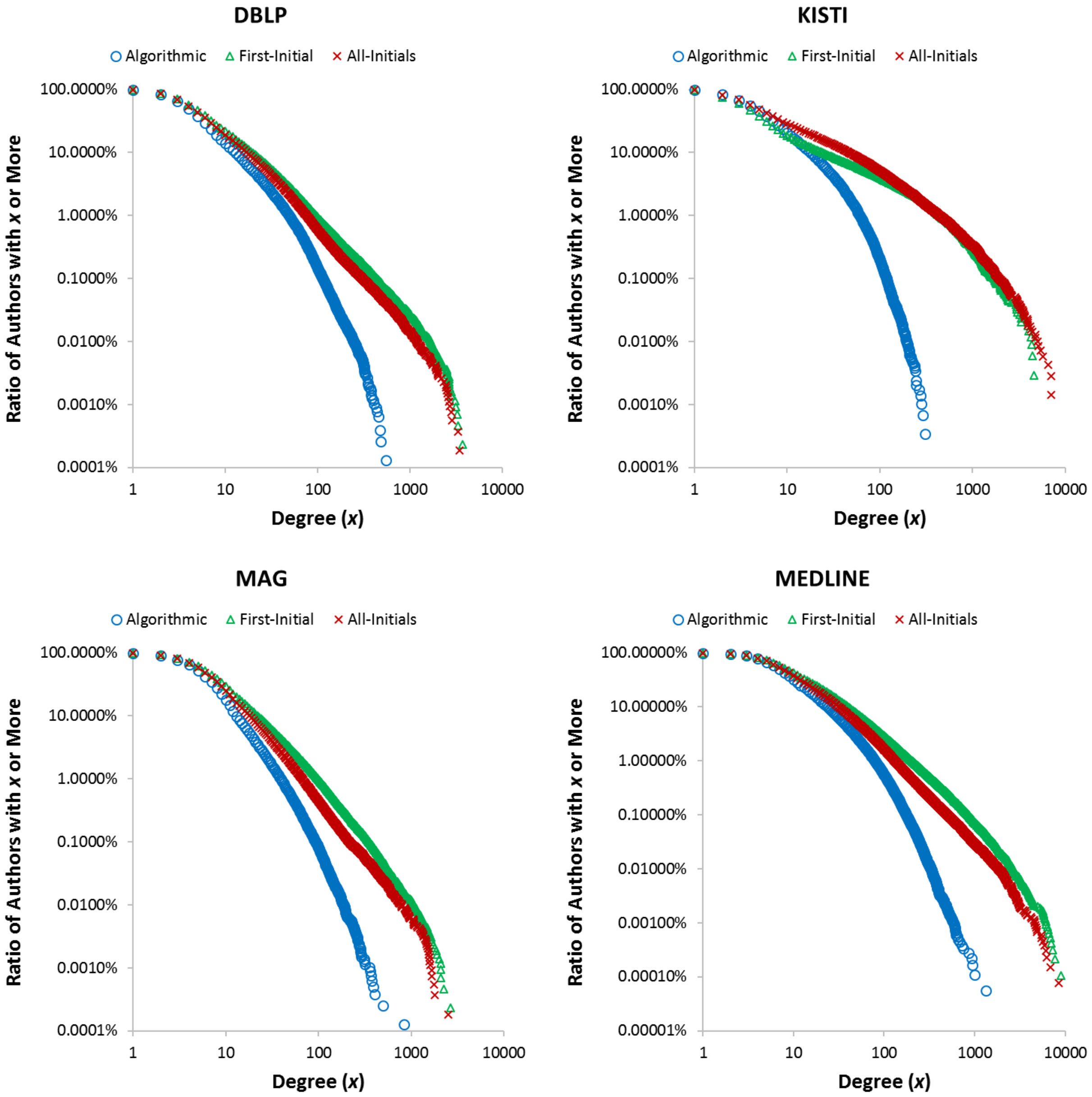


Figure 9: Cumulative Log-Log Plot of Vertex Degree Distribution

Third, the distribution plots from the coauthorship networks disambiguated by initial-based disambiguation show straighter trendlines than those from algorithmically disambiguated networks, which seems to fit a power-law slope. To better find the power-law fit, scholars have used three different approaches. Some scholars have visually compared a degree distribution with a straight line that covers the largest portion of the distribution plot (e.g., Milojević, 2010) and have calculated its fit using R-squared value (e.g., Barabási et al., 2002; Newman, 2001b; Perc, 2010). This method has been criticized for finding power-fit slopes against any plot, which are in most cases false positives (Clauset et al., 2009; Stumpf & Porter, 2012).

Others have turned to a more rigorous measure proposed by Clauset et al. (2009), which uses a "maximum-likelihood fitting method with a goodness-of-fit test based on the Kolmogorov-Smirnov statistic." According to this strict measure, many power-law distributions in previous network studies turned out to be false positives (Clauset et al., 2009). The plots in Figure 9 were tested for power-law fitting using this improved measure. The result is summarized in Table 11. Power-law fitting was found (*p*-value > .10) against plots generated by algorithmic disambiguation for DBLP, KISTI, and MEDLINE. The power-law regime describes at best 1.8% of all authors, which makes the fitting results useless (Stumpf & Porter, 2012). All-initials method produced one case of power-law fitting plot for MAG, but with a very low number of authors (1.6%) described.

Recently, a few scholars introduced a measure where a network's degree distribution is compared to that of a network that is synthetically created with a similar number of vertices, average degree, and number of unique edges, but shows a power-law distribution in degree (Kim & Diesner, 2015; Kim et al., 2016). This thesis uses the third approach for checking whether a power-law distribution can be a plausible description of the degree distribution in a network.

Table 11: Result of Power-Law Fitting Using a Statistical Measure

Disambiguation Method	**Parameters**	**Data**			
		DBLP	**KISTI**	**MAG**	**MEDLINE**
Algorithmic	x-min (coverage)	81 **(0.3%)**	103 **(1.8%)**	100 **(0.8%)**	137 **(0.2%)**
	slope	4.10	5.42	4.43	4.40
	p-value	0.99	0.35	0.57	0
	GOF	0.0077	0.0209	0.0173	0.0166
First-Initial	x-min (coverage)	53 **(2.6%)**	8 **(22.9%)**	111 **(0.8%)**	101 **(2.9%)**
	slope	2.57	1.74	2.87	2.53
	p-value	0.08	0	0.07	0
	GOF	0.0069	0.0292	0.0120	0.0081
All-Initials	x-min (coverage)	45 **(2.5%)**	4 **(57.6%)**	53 **(1.6%)**	81 **(2.4%)**
	slope	2.72	1.72	2.93	2.77
	p-value	0	0	0.23	0
	GOF	0.0112	0.0266	0.0069	0.0070

For each dataset, three synthetic networks generated by the preferential attachment proposed in Barabási et al. (2002) were created using *Pajek*. They had "the same or similar number of unique authors, edges, and average degrees as the networks resulting from algorithmic, first-initial, and all-initials disambiguation" (Kim & Diesner, 2015). Especially, they are also called "scale-free" networks as their vertex degree distributions follow a power-law slope over any range (i.e., scale) of *x* values. The properties of simulated (synthetic) networks are summarized in Table 12. Cumulative log-log plots of vertex degree distributions from simulated networks are shown in Figure 10 (black circles, triangles, or crosses per disambiguation method), along with those by algorithmic (blue circles), first-initial (green triangles), and all-initials (red crosses) methods.

Table 12: Summary of Empirical and Synthetic Networks per Disambiguation Method

Disambiguation Method	**Type & Parameters**	**Data**							
		DBLP		**KISTI**		**MAG**		**MEDLINE**	
	Network Type	Empirical	Synthetic	Empirical	Synthetic	Empirical	Synthetic	Empirical	Synthetic
Algorithmic	No. of Authors	777,882	777,580	291,890	291,559	794,212	793,935	1,816,093	1,815,538
	No. of Edges	2,394,946	2,390,940	1,129,044	1,125,019	2,667,998	2,663,276	10,061,993	10,048,301
	Avg. Degree	6.16	6.15	7.74	7.72	6.72	6.71	11.08	11.07
First-Initial	No. of Authors	437,215	437,126	34,040	33,986	428,859	428,112	950,409	950,409
	No. of Edges	2,231,050	2,223,911	385,080	374,333	2,437,562	2,428,285	9,751,972	9,715,924
	Avg. Degree	10.21	10.18	22.63	22.03	11.37	11.34	20.52	20.45
All-Initials	No. of Authors	536,959	535,634	70,347	70,313	545,518	545,509	1,322,674	1,322,674
	No. of Edges	2,312,157	2,306,253	990,053	967,769	2,495,808	2,488,860	10,257,476	10,224,209
	Avg. Degree	8.61	8.61	28.15	27.53	9.15	9.12	15.51	15.46

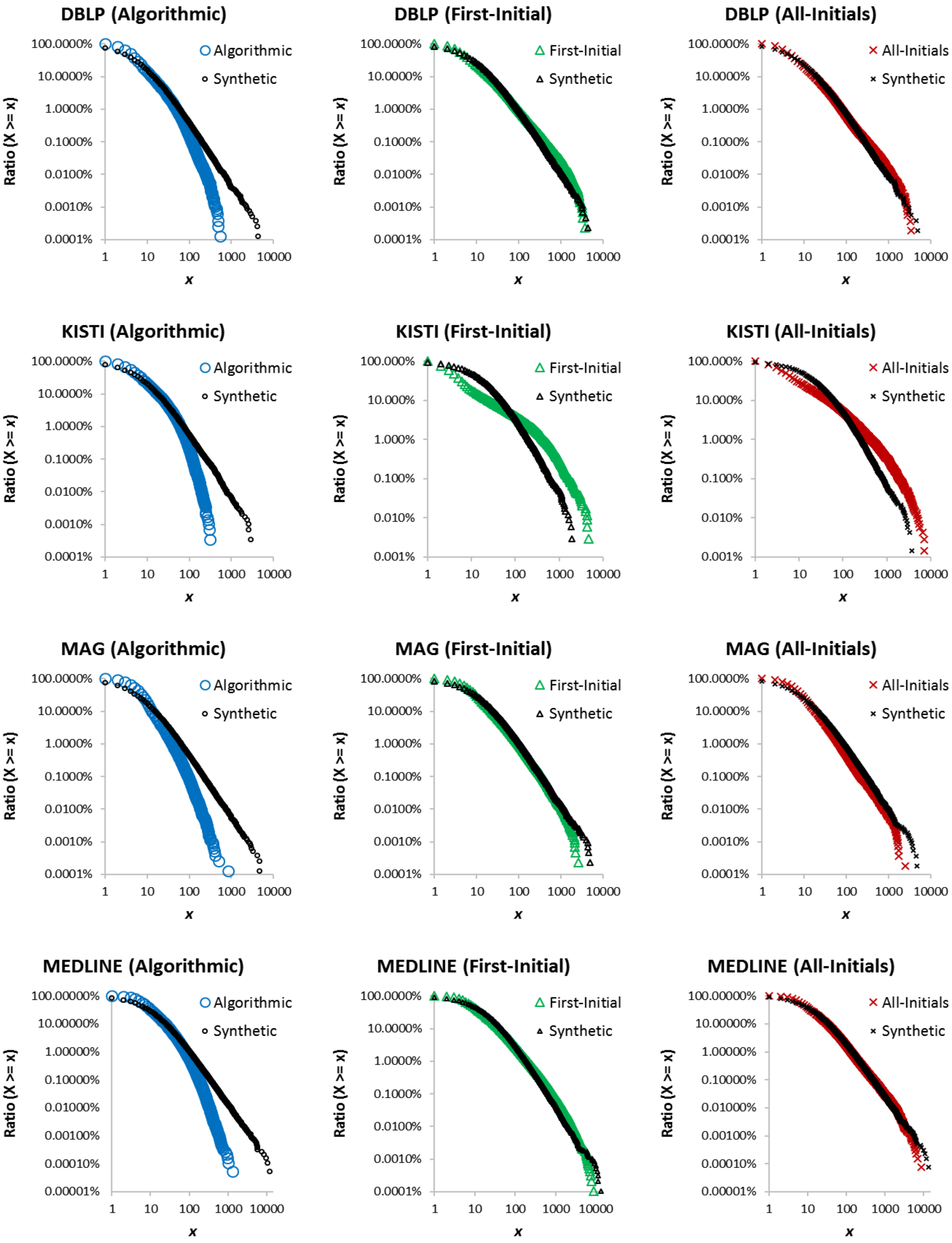

DBLP (Algorithmic)
DBLP (First-Initial)
DBLP (All-Initials)
KISTI (Algorithmic)
KISTI (First-Initial)
KISTI (All-Initials)
MAG (Algorithmic)
MAG (First-Initial)
MAG (All-Initials)
MEDLINE (Algorithmic)
MEDLINE (First-Initial)
MEDLINE (All-Initials)

Figure 10: Cumulative Log-Log Plot of Degree Distribution Compared to Scale-Free Degree Distribution

In Figure 10, if a degree distribution of a coauthorship network fits into a power-law distribution, it should align with the distribution plot generated by a synthetic network. The figure suggests that overall, plots from algorithmic disambiguation show departure from the ideal power-law degree distribution. In contrast, initial-based disambiguation produced degree plots that seem close to the slopes of their ideal counterparts across many x values except for KISTI. This illustrates that depending on the choices of name disambiguation methods, different or same topologies can characterize the same network data.

5.9 CENTRALIZATION AND DENSITY

Like the Gini coefficient for production, degree centralization can be used to measure the concentration of degree in a network. Figure 11 shows the temporal change in the degree centralization in four datasets per disambiguation method: algorithmic disambiguation (circles), first-initial method (triangles), and all-initials based one (crosses). The centralization is marked up to the target year for each disambiguation method. An inset figure shows the ratio of error calculated as follows.

$$Error\ Ratio\ (\%) = \frac{Val_A - Val_B}{Val_B} \times 100$$

Here, Val_A refers to the centralization calculated for data disambiguated by first-initial or all-initials method, while Val_B refers to the centralization obtained from algorithmically disambiguated data.

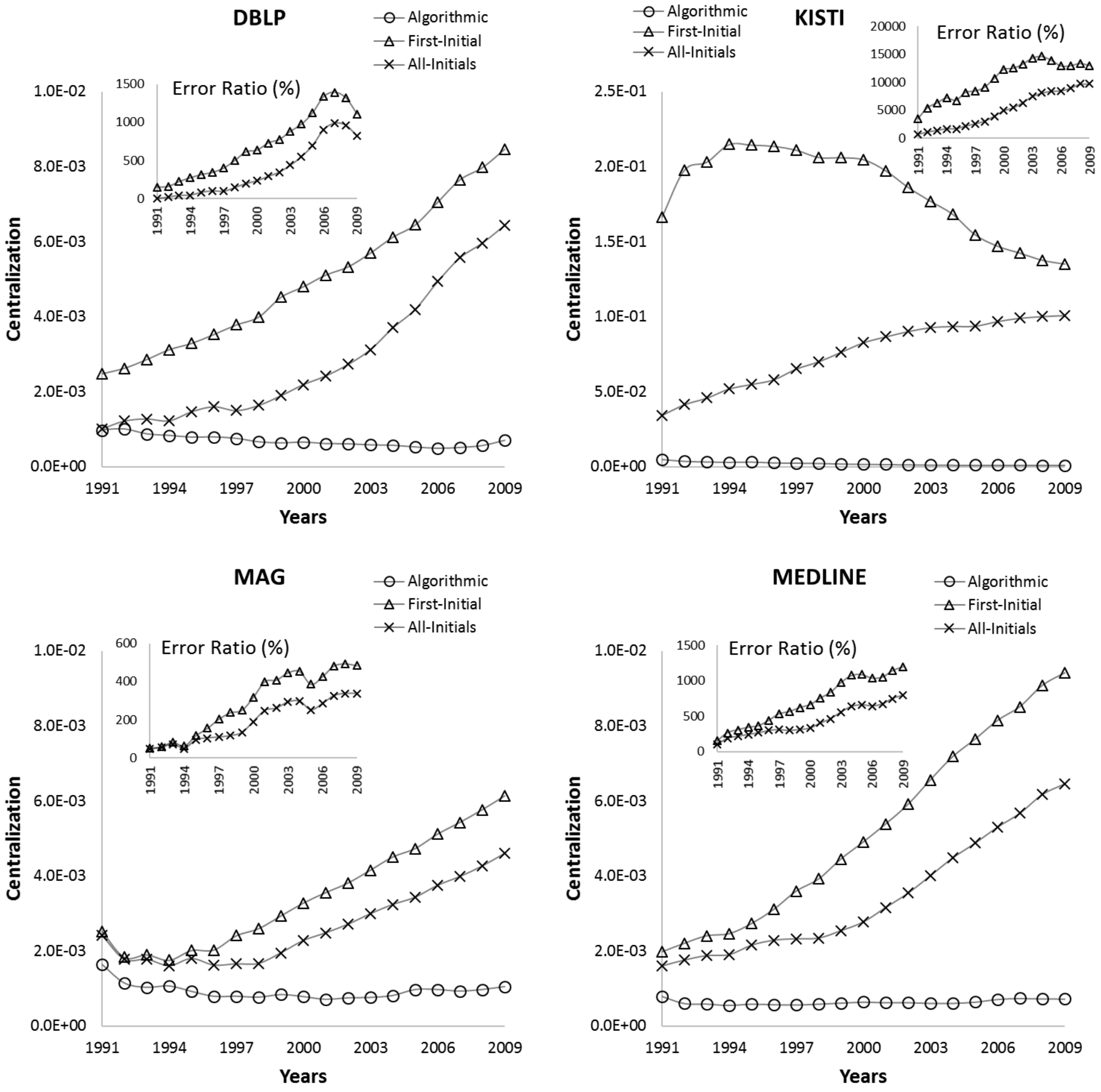


Figure 11: Trend of Degree Centralization

The gap of changes among disambiguation methods gets wider over time. In particular, the error ratios exceed several hundred percent in all datasets. This tendency can be explained by the increased average degree by initial-based disambiguation. The inflated number of high-degree authors facilitates the concentration of degree, leading to high centralization. The dramatic change is, however, due to the calculation scheme of network centralization. The denominator has the theoretical maximum differences between pairwise vertices. During this process, the decreased number of vertices disproportionally reduces the value of the denominator, which pushes up the centralization value with the increased numerator by high-degree authors.

Similarly, network density trends show substantial change in Figure 12 but in a different direction: while centrality increased over time, density decreased. The widening gaps between initial-based and algorithmically disambiguated plots are mainly due to the denominator in the density equation. As the number of vertices decreases from A to B, the denominator also decreases by the difference between A(A-1)/2 and B(B-1)/2. In contrast, the decreasing trend of all plots over time is due to the increase of vertices per disambiguation method.

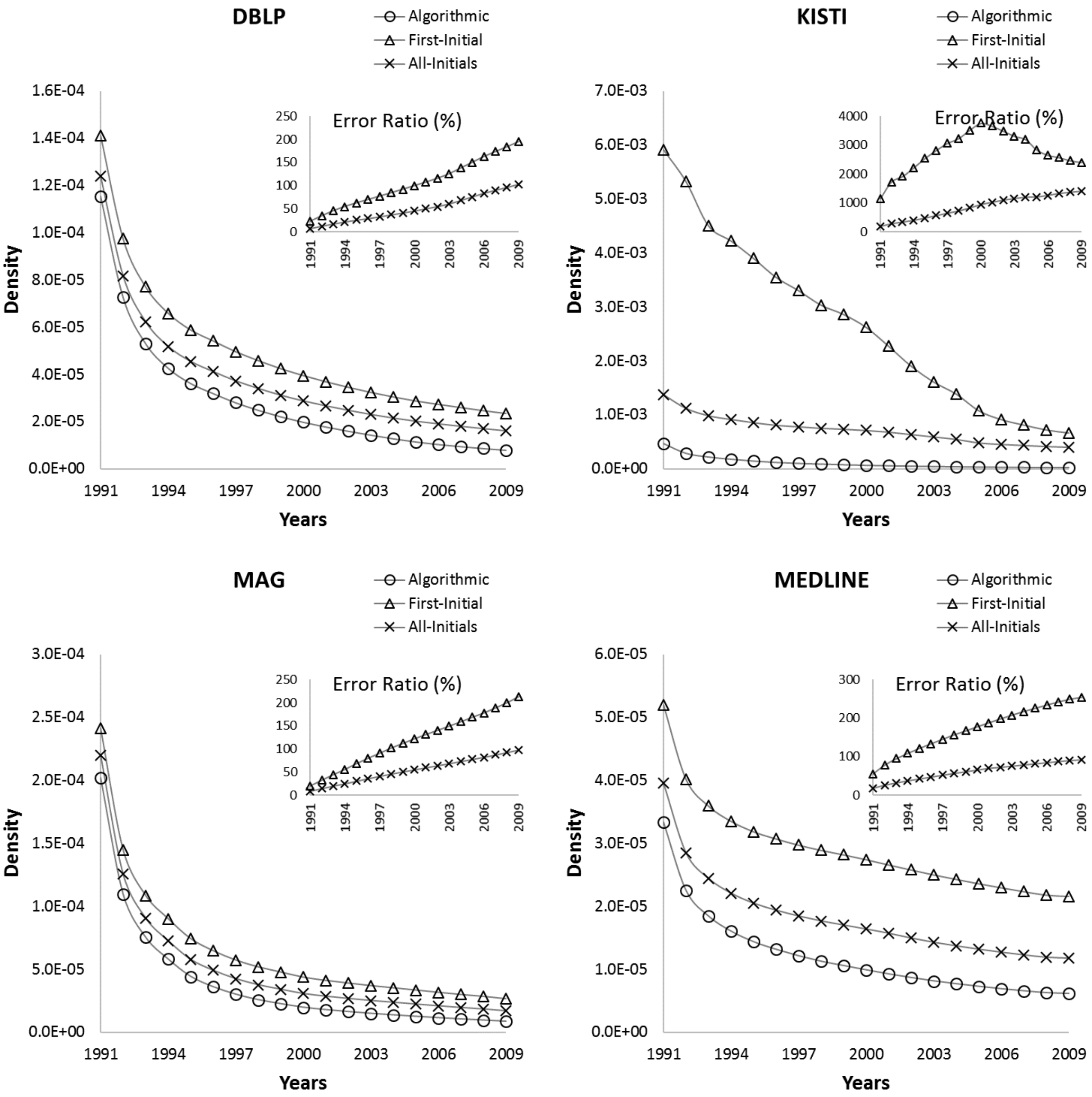


Figure 12: Trend of Network Density

5.10 RATIO OF THE LARGEST COMPONENT

Figure 13 shows the ratios of the largest components over time in four datasets per disambiguation method: algorithmic disambiguation (circles), first-initial method (triangles), and all-initials based one (crosses). The ratio is marked up to the target year for each disambiguation method. An inset figure shows the ratio of error calculated as follows.

$$Error\ Ratio\ (\%p) =\ Ratio_A - Ratio_B$$

Here, $Ratio_A$ refers to the ratio (%) of the largest component calculated for data disambiguated by first-initial or all-initials method, while $Ratio_B$ refers to the ratio of the largest component obtained from algorithmically disambiguated data.

Overall, the ratios of the largest components of algorithmically disambiguated networks are positioned below those for networks disambiguated by initial-based disambiguation methods. This shows that coauthorship networks after initial-based author name disambiguation “tend to inflate the ratios of the largest components in comparison to those of algorithmically disambiguated networks” (Kim &Diesner, 2015). The merging of author identities can account for this. “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” (Kim & Diesner, 2015).

“The gap of ratios between algorithmically disambiguated data and initial-based processed data increased for some time and then moderately decreased. 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 in the component size would not be noticeable compared to the situation when they are in separate components or isolated from components before merging” (Kim & Diesner, 2015). At the early years, “many of the 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" (Kim & Diesner, 2016).

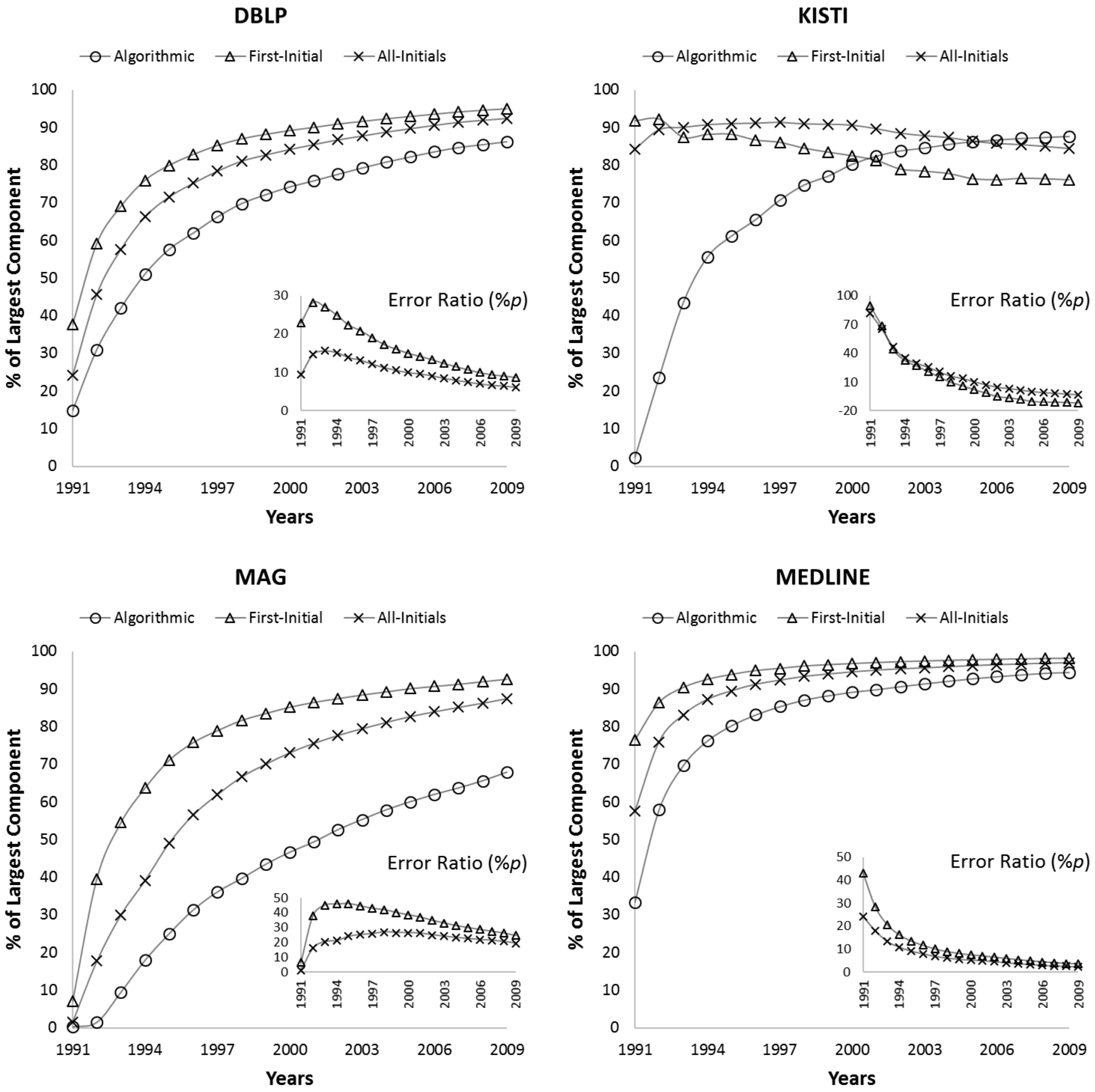


Figure 13: Trend of Ratio of the Largest Component

5.11 AVERAGE SHORTEST PATH LENGTHS

The average shortest path lengths decrease over time regardless of author name disambiguation method in Figure 14. Such a decreasing trend was also found in previous coauthorship network evolution research (e.g., Franceschet, 2011; Martin et al., 2013; Perc, 2010). In the figure, specifically, the average shortest path lengths of the network disambiguated by advanced algorithms are larger than those of networks disambiguated by first-initial and all-initials methods. This can be explained in conjunction with increased sizes of the largest components. As more authors were attached to larger components over time, merged authors in networks "act as bridges connecting authors who were unreachable, or by providing shorter paths for authors who were reachable with longer paths," thereby decreasing the shortest path lengths among authors (Kim & Diesner, 2015, 2016).

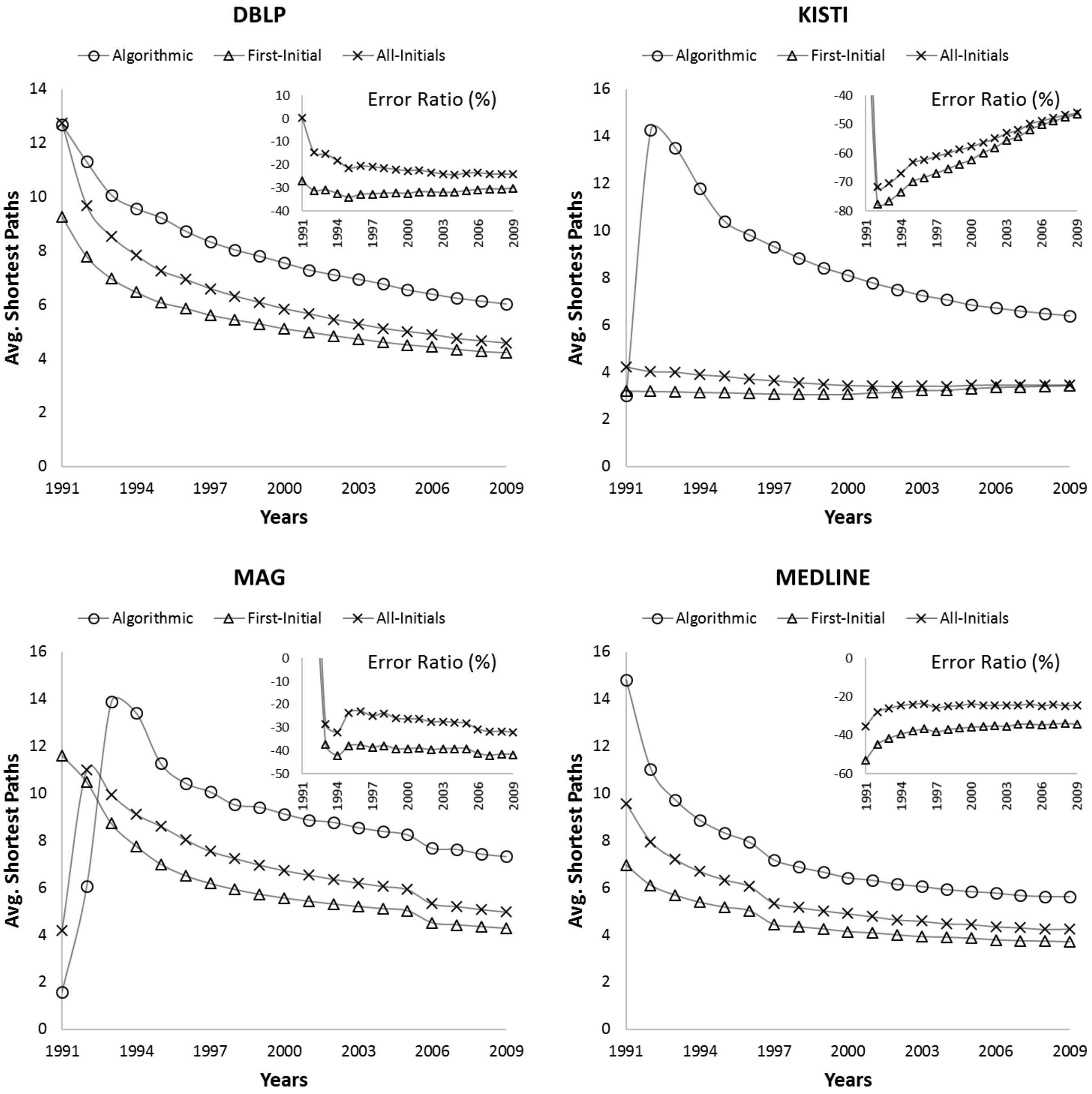


Figure 14: Trend of Average Shortest Path Length

5.12 DEGREE ASSORTATIVITY

Degree assortativity attempts to characterize the pattern of edge formation in a network using the extent to which vertices in the network show a tendency to be linked to similar others in terms of vertex degree. Figure 15 shows the temporal change of the degree assortativity in four datasets per disambiguation method: algorithmic disambiguation (circles), first-initial method (triangles), and all-initials based one (crosses). The assortativity is marked up to the target year for each disambiguation method. An inset figure shows the ratio of error calculated as follows.

$$Error\ Ratio\ (\%) = \frac{Val_A - Val_B}{Val_B} \times 100$$

Here, Val_A refers to the assortativity calculated for data disambiguated by first-initial or all-initials method, while Val_B refers to the assortativity obtained from algorithmically disambiguated data.

The overall trend in DBLP, MAG, and MEDLINE shows that authors tend to collaborate with other authors who have similar numbers of coauthors, and such a tendency decreases over time, reaching a low level of 0.10 or less in most datasets regardless of disambiguation methods. Unlike other measures so far, however, the differences by disambiguation methods do not seem to show any consistent patterns. A noticeable trend happened for KISTI, where initial-based disambiguation methods produce a degree assortativity below zero: authors with a high degree centrality appear to prefer to collaborate with others with a low degree centrality.

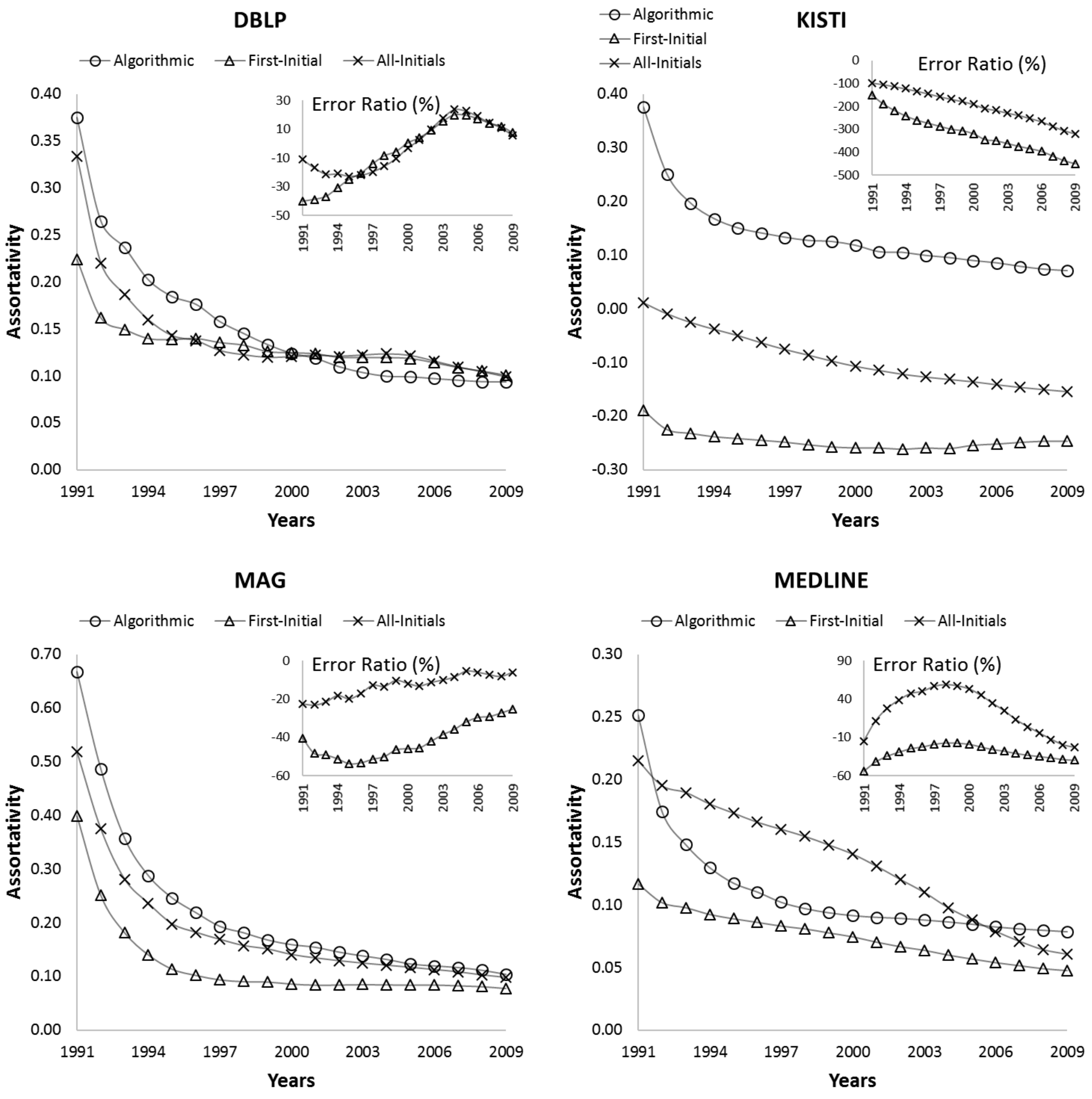


Figure 15: Trend of Degree Assortativity

5.13 TRANSITIVITY

Like degree assortativity, transitivity attempts to characterize the patterns of edge formation in a network based on the shared vertices between a pair of vertices: if two vertices share a common vertex (or vertices), they are likely to form an edge with each other. Figure 16 shows the temporal change of the transitivity in four datasets per disambiguation method: algorithmic disambiguation (circles), first-initial method (triangles), and all-initials based one (crosses). The transitivity is marked up to the target year for each disambiguation method. An inset figure shows the ratio of error calculated as follows.

$$Error\ Ratio\ (\%) = \frac{Val_A - Val_B}{Val_B} \times 100$$

Here, Val_A refers to the transitivity calculated for data disambiguated by first-initial or all-initials method, while Val_B refers to the transitivity obtained from algorithmically disambiguated data.

Transitivity of networks shows an overall decreasing trend across all datasets. This implies that authors in each dataset become increasingly “less inclined to collaborate with others” when sharing one or more common coauthors (Kim & Diesner, 2016). This also indicates that, contrary to several edge (link) prediction studies assuming transitivity as a strong predictor (e.g., Guns & Rousseau, 2014; Liben-Nowell & Kleinberg, 2007), sharing coauthors does not seem to be a strong predictor of coauthorship edge formation because the decreasing transitivity means scholars tend not to form edges even if they share one or more collaborators.

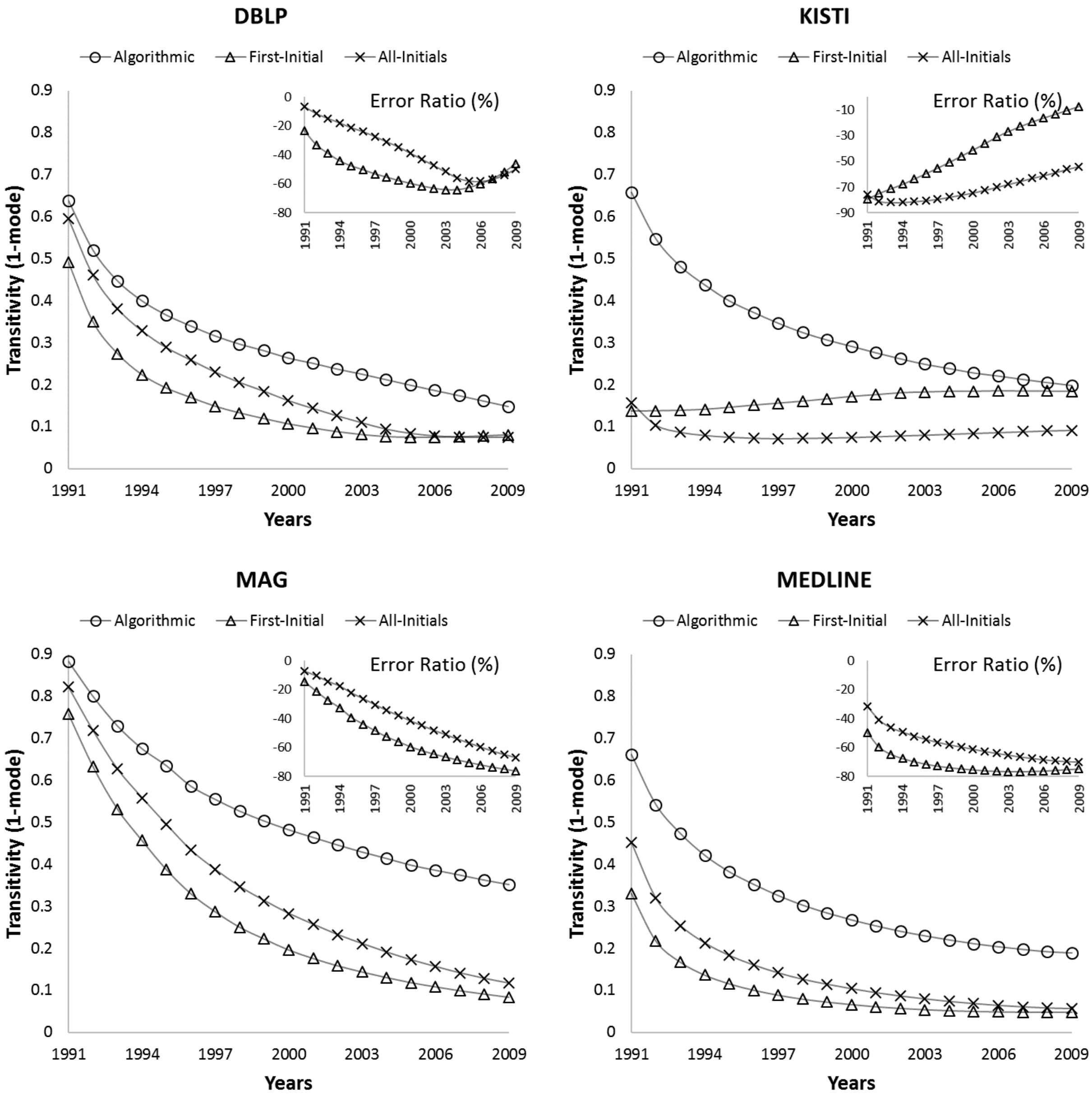


Figure 16: Trend of Transitivity for One-Mode Network

The gaps between transitivity trendlines that are disambiguated by initial-based and algorithmic methods increase over time. Transitivity 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 collaborated, a triangle fails to form. If this happens often when authors are merged, the clustering coefficient of the network begins to decrease" (Kim &Diesner, 2015). This transitivity deflation indicates that initial-based disambiguation makes a scholar appear to be more reluctant to collaborate with others who once worked with a common coauthor than they actually are.

In comparison to transitivity calculated for one-mode networks, Figure 17 shows the situation when transitivity is calculated for two-mode networks. The two-mode transitivity hovers below 0.30 in DBLP, MAG, and MEDLINE. An interesting point is that two-mode network transitivity seems to be quite stable (despite moderate decreases for MAG and MEDLINE) for algorithmically disambiguated DBLP and KISTI networks, while initial-based disambiguation produced quite confusing trends. This is contrasted to the observation that one-mode transitivity shows a similar trend for all three disambiguation methods over time. This indicates that two-mode network transitivity may better capture the true differences in network clustering patterns between disambiguation methods than one-mode network transitivity does.

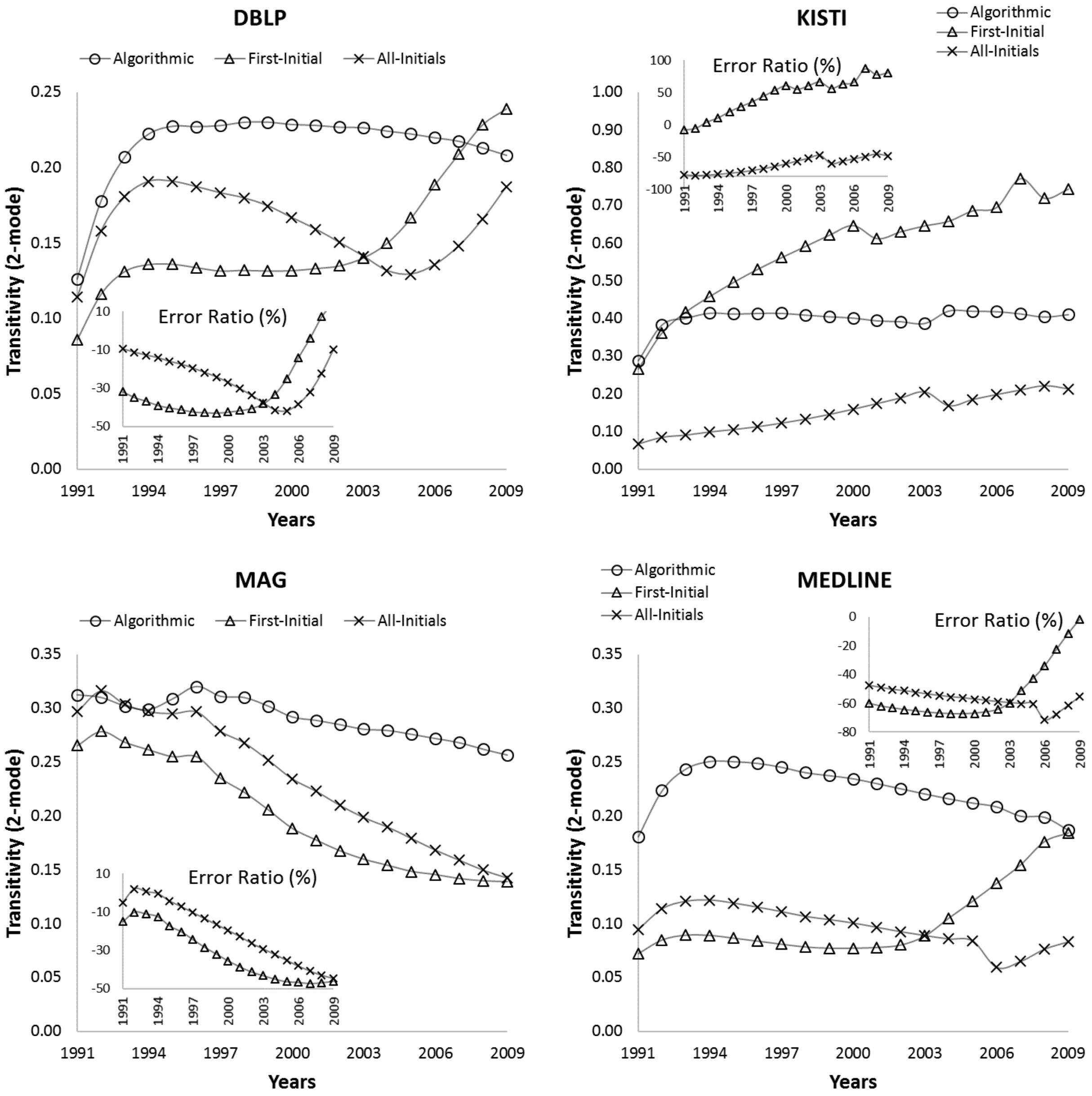


Figure 17: Trend of Transitivity for Two-Mode Network

5.14 *K*-2-PATHS

In coauthorship networks, *k*-2-paths represent situations where two authors do not collaborate with each other even when they share coauthors. Table 13 reports the frequencies of *k*-2-paths (*k*

= 1…15) per disambiguation method in four empirical datasets. A note is that, due to the computational complexity, subsets of DBLP (608,990 papers), KISTI (215,410), MAG (221,755), and MEDLINE (386,862) that cover publications between 2006 and 2009 were used for calculating *k*-2-paths hereafter.

Disambiguation methods were found to affect the observed frequencies of *k*-2-paths. Table 13 shows the change of frequencies of *k*-2-paths (*k* = 1…15) by initial-disambiguation compared to those by algorithmic disambiguation. In the table, the frequency change by first-initial and all-initials methods are reported by what factor (i.e., a factor of 10) the counts by initial-based methods are larger than those by algorithmic disambiguation.

As the *k* increases, the gaps between the frequencies of *k*-2-paths by initial-based methods and algorithmic disambiguation become larger. For example, in DBLP, the frequency of *k* = 1 by algorithmic disambiguation increased by 3.5 times via first-initial disambiguation and by 2.7 times via all-initials disambiguation. For *k* = 15, however, the gaps reached 2,820 times by first-initial method and 1,571 by all-initials. The merging effect illustrated in Figure 2 can explain this finding. As multiple authors are merged into one, their coauthors become embedded into *k*-2-paths. If two authors happen to have many coauthors whose identities are merged, they become embedded into a high order of *k*-2-paths. As first-initial method introduces more merging than the all-initials method, the frequencies of *k*-2-paths via first-initial method increased at a higher rate than those pre-processed via all-initials method.

Table 13: Change of Numbers of k-2-Paths by Initial-Based Methods

k	**DBLP**			**KISTI**			**MAG**			**MEDLINE**		
	Algorithmic (Count)	First-Initial (Times)	All-Initials (Times)	Algorithmic (Count)	First-Initial (Times)	All-Initials (Times)	Algorithmic (Count)	First-Initial (Times)	All-Initials (Times)	Algorithmic (Count)	First-Initial (Times)	All-Initials (Times)
1	35.2M	3.5	2.7	13.1M	1.1	7.2	17.7M	6.7	3.9	12.3M	12.1	8.5
2	3.2M	4.8	3.4	2.2M	2.9	12.9	2.7M	2.9	1.8	2.0M	7.0	4.1
3	680,506	8.1	5.5	739,132	4.7	18.7	702,060	3.3	2.0	560,035	7.9	4.1
4	179,107	15.1	10.0	304,199	6.9	26.6	248,020	4.0	2.5	211,491	9.5	4.6
5	54,958	28.7	18.5	137,863	9.8	38.8	99,790	5.4	3.3	85,655	13.1	6.0
6	19,326	53.1	33.6	64,559	14.7	59.1	42,535	7.6	4.6	36,202	19.4	8.4
7	7,383	97.6	60.6	31,701	21.9	90.2	19,039	11.2	6.7	14,594	32.8	13.8
8	3,172	168.5	102.6	16,283	33.5	136.3	8,394	17.9	10.6	5,346	65.0	26.5
9	1,530	268.6	161.5	8,899	49.4	199.3	4,089	26.9	15.9	1,625	162.4	64.1
10	776	422.6	249.4	5,146	71.1	280.7	1,884	44.0	25.5	466	442.8	171.6
11	424	625.5	366.6	3,097	99.2	386.0	949	68.4	39.6	139	1,193.4	452.1
12	232	953.0	549.6	1,881	141.1	535.5	532	97.8	55.6	33	4,158.1	1,528.1
13	133	1,405.6	795.4	1,120	206.7	766.7	309	135.3	76.8	24	4,754.0	1,692.5
14	87	1,835.6	1,029.2	786	252.4	939.6	206	169.8	95.1	8	12,108.4	4,259.1
15	49	2,820.8	1,571.2	470	378.0	1,360.9	119	244.2	136.1	4	20,780.5	7,112.8

This observation implies that depending on the choice of disambiguation method scholars may be motivated to conduct coauthorship network research in different directions. According to network theorists, two network actors who share common neighbors tend to form edges (Holland & Leinhardt, 1970), which has been a theoretical basis of studying clustering (transitivity) for understanding network evolution and predicting edge formation. From the perspective of this research trend, the *k*-2-paths may be abnormal because they show the failure of edge formation between dyads that share common vertices. The dramatic increase of *k*-2-paths, especially with high *k* values, by initial based disambiguation can lead scholars to formulate research questions on why authors do not collaborate with each other even when they share many coauthors. Like the case of transitivity in one-mode networks (see 5.13 Transitivity), this also implies that studies relying on shared vertices to predict edge formation are likely to find their predictor perform poor for coauthorship networks, especially disambiguated by name initials.

5.15 IMPACT OF NAME AMBIGUITY COMPARED TO RANDOM NETWORKS

Scholars have attempted to infer whether observed (or empirical) networks show patterns that are not likely to be found in random networks. For this, they have simulated random networks containing the same or similar number of vertices, edges, or degree distribution and compared network properties of those random networks to those of empirical ones to see the differences in frequencies of specific patterns, configurations, or topologies.

Table 14 and Table 15 report results of comparing network properties of empirical networks per disambiguation method for each dataset and those of random networks that have the same or similar numbers of vertices and edges. For each disambiguation method in a dataset, 100 Erdős–Rényi random networks were generated and six network metrics – centralization, ratio of the largest component, average shortest paths, assortativity, one-mode transitivity, and two-mode transitivity -- were calculated on each random network. A note is that the numbers of vertices and edges, average degree, and density were the same or similar with negligible errors between empirical and random networks. Then, the metric values were averaged and standard deviation is reported. The same metrics were also calculated on each empirical network per disambiguation method. A change ratio is reported to show how different random networks are when compared to empirical networks in terms of a metric. The C-Ratio shows the ratio of difference calculated as follows.

$$C - Ratio\ (\%) = \frac{Val_A - Val_B}{Val_B} \times 100$$

Here, Val_A represents the metric value calculated for empirical networks, while Val_B refers to the value obtained from random networks. For the Ratio of the Largest Component, C-Ratio is calculated simply as $Val_A - Val_B$ in percentage points.

Table 14: Comparison of Empirical Versus Random Network Per Measure (DBLP and KISTI)

Data	Disambiguation Method	Network Type	Centralization	% of Largest Component	Avg. Shortest Paths	Assortativity	Transitivity (1-Mode)	Transitivity (2-Mode)
DBLP	Algorithmic	Random (SD)	0.000020 (0.000002)	99.79 (0.005327)	7.68 (0.02)	-0.000010 (0.000620)	0.000008 (0.000001)	0.000022 (0.000002)
		Empirical (C-Ratio)	0.000700 (3,472%)	86.23 (-14%*p*)	6.03 (-22%)	0.093554 (-910,026%)	0.148549 (1,802,965%)	0.208098 (930,716%)
	First-Initial	Random (SD)	0.000041 (0.000003)	100.00 (0.000856)	5.84 (0.02)	0.000040 (0.000667)	0.000023 (0.000001)	0.000062 (0.000002)
		Empirical (C-Ratio)	0.008469 (20,454%)	94.96 (-5%*p*)	4.21 (-28%)	0.100650 (253,151%)	0.080119 (342,053%)	0.239042 (382,748%)
	All-Initials	Random (SD)	0.000032 (0.000002)	99.98 (0.001837)	6.37 (0.02)	-0.000063 (0.000625)	0.000016 (0.000001)	0.000044 (0.000002)
		Empirical (C-Ratio)	0.006439 (20,312%)	92.39 (-8%*p*)	4.57 (-28%)	0.098982 (-157,813%)	0.074710 (465,557%)	0.187331 (426,399%)
KISTI	Algorithmic	Random (SD)	0.000055 (0.000004)	99.96 (0.003987)	6.40 (0.02)	0.000151 (0.001023)	0.000027 (0.000003)	0.000079 (0.000005)
		Empirical (C-Ratio)	0.001029 (1,780%)	87.63 (-12%*p*)	6.39 (-0.17%)	0.070306 (46,311%)	0.198643 (738,155%)	0.410386 (520,783%)
	First-Initial	Random (SD)	0.000645 (0.000041)	100.00 (-)	3.69 (0.01)	-0.000061 (0.001704)	0.000666 (0.000015)	0.006051 (0.000016)
		Empirical (C-Ratio)	0.134864 (20,808%)	75.96 (-24%*p*)	3.42 (-7%)	-0.246511 (407,335%)	0.184341 (27,579%)	0.742752 (12,175%)
	All-Initials	Random (SD)	0.000364 (0.000026)	100.00 (-)	3.71 (0.01)	0.000020 (0.000914)	0.000400 (0.000006)	0.001308 (0.000010)
		Empirical (C-Ratio)	0.100760 (27,617%)	84.41 (-16%*p*)	3.45 (-7%)	-0.154002 (-752,033%)	0.091272 (22,739%)	0.212154 (16,122%)

Table 15: Comparison of Empirical Versus Random Network Per Measure (MAG and MEDLINE)

Data	Disambiguation Method	Network Type	Centralization	% of Largest Component	Avg. Shortest Paths	Assortativity	Transitivity (1-Mode)	Transitivity (2-Mode)
MAG	Algorithmic	Random	0.000020	99.88	7.35	-0.000081	0.000008	0.000024
		(SD)	(0.000001)	(0.003928)	(0.02)	(0.000648)	(0.000001)	(0.000003)
		Empirical	0.001050	67.82	7.32	0.103953	0.352056	0.256649
		(C-Ratio)	(5,166%)	(-32%*p*)	(-0.42%)	(-128,441%)	(4,269,347%)	(1,089,959%)
	First-Initial	Random	0.000045	100.00	5.61	-0.000069	0.000027	0.000057
		(SD)	(0.000003)	(0.000544)	(0.02)	(0.000707)	(0.000002)	(0.000003)
		Empirical	0.006132	92.58	4.27	0.077938	0.083516	0.138600
		(C-Ratio)	(13,637%)	(-7%*p*)	(-24%)	(-112,989%)	(314,117%)	(243,011%)
	All-Initials	Random	0.000032	99.99	6.21	-0.000018	0.000017	0.000040
		(SD)	(0.000002)	(0.001451)	(0.02)	(0.000657)	(0.000001)	(0.000003)
		Empirical	0.004595	87.39	4.97	0.097769	0.116433	0.142551
		(C-Ratio)	(14,291%)	(-13%*p*)	(-20%)	(-541,076%)	(699,273%)	(358,812%)
MEDLINE	Algorithmic	Random	0.000011	100.00	6.25	0.000044	0.000006	0.000013
		(SD)	(0.000001)	(0.000316)	(0.02)	(0.000314)	(0.000000)	(0.000001)
		Empirical	0.000723	94.41	5.63	0.078527	0.188860	0.186844
		(C-Ratio)	(6,415%)	(-6%*p*)	(-10%)	(179,368%)	(3,107,201%)	(1,386,748%)
	First-Initial	Random	0.000027	100.00	4.85	-0.000041	0.000022	0.000035
		(SD)	(0.000002)	(-)	(0.01)	(0.000330)	(0.000001)	(0.000002)
		Empirical	0.009406	98.07	3.70	0.047581	0.047587	0.183970
		(C-Ratio)	(35,042%)	(-2%*p*)	(-24%)	(-115,273%)	(220,049%)	(519,467%)
	All-Initials	Random	0.000017	100.00	5.45	-0.000067	0.000012	0.000021
		(SD)	(0.000001)	(0.000034)	(0.02)	(0.000306)	(0.000000)	(0.000001)
		Empirical	0.006449	96.88	4.26	0.060316	0.056830	0.083038
		(C-Ratio)	(36,793%)	(-3%*p*)	(-22%)	(-90,538%)	(486,274%)	(389,827%)

Overall, the ratio of the largest component and average shortest path lengths showed the lower differences between empirical and random networks than other measures. The empirical networks showed C-Ratios of -32%*p* ~ 0%*p* for the ratio of the largest component and -28% ~ -0.17% for average shortest paths when compared to random networks. Other measures, especially transitivity (both one-mode and two-mode versions), showed a dramatic change from random networks. While three measures – centralization and two transitivity metrics – showed an increase from random networks, two measures—such as the ratio of the largest component and average shortest path lengths—showed decreases. Assortativity showed a mix of increase and decrease.

The observations indicate that, regardless of disambiguation method, the empirical networks are different from random networks. A specific disambiguation method may not affect statistical inference of properties of a network. The name ambiguity can, however, affect the level to which an empirical network shows a tendency toward a specific topology (e.g., Small-Worldness) or a local pattern (e.g., transitivity). For example, algorithmic disambiguation consistently showed a lower C-Ratio for average shortest paths than initial-based disambiguation against corresponding random networks. As the Small-Worldness is parsimoniously defined as networks having average shortest path lengths close to those of random networks and averaged local clustering coefficient (averaged ego network's densities) larger than those of random networks, it can be conjectured that, given the same averaged local clustering coefficient, algorithmically disambiguated networks would exhibit a stronger tendency toward the Small-Worldness than networks disambiguated by name initials.

Meanwhile, Table 16 and Table 17 report the frequencies of *k*-2-paths ($k = 1…15$) per disambiguation method in four empirical datasets and those found in two-mode random networks corresponding to each of the empirical datasets. The two-mode random networks for k-2-paths calculation were generated once per disambiguation method due to (1) computational complexity and (2) marginal differences (less than 5%) among 100 simulated random networks for the algorithmically disambiguated DBLP network. Regardless of disambiguation methods, networks from empirical datasets contain larger numbers of *k*-2-paths across all *k* values than randomly generated networks. In random networks, *k*-2-paths were found within *k* = 1~ 3 for DBLP, MAG, and MEDLINE, and within *k* = 1 ~ 9 for KISTI. This observation might lead us to conclude that empirical networks, whether they are disambiguated by algorithms or name initials, show a higher tendency of generating more *k*-2-paths, especially with high *k* values, than random networks.

Table 16: Frequencies of K-2-Paths in Empirical Versus Random Networks (DBLP and KISTI)

k	**DBLP**						**KISTI**					
	Algorithmic		**First-Initial**		**All-Initials**		**Algorithmic**		**First-Initial**		**All-Initials**	
	Empirical	Random	Empirical	Random	Empirical	Random	Empirical	Random	Empirical	Random	Empirical	Random
1	35.2M	69.7M	158.4M	123.2M	131.8M	100.7M	13.1M	42.4M	27.0M	185.3M	107.6M	161.5M
2	3.2M	12,237	18.8M	101,444	14.4M	47,404	2.2M	30,260	8.7M	56.1M	31.2M	6.2M
3	680,506	3	6.2M	66	4.4M	27	739,132	17	4.2M	11.8M	14.6M	175,020
4	179,107	0	2.9M	0	2.0M	0	304,199	0	2.4M	2.0M	8.4M	3,981
5	54,958	0	1.6M	0	1.1M	0	137,863	0	1.5M	271,075	5.5M	83
6	19,326	0	1.0M	0	669,302	0	64,559	0	1.0M	32,470	3.9M	1
7	7,383	0	728,115	0	455,097	0	31,701	0	727,081	3,380	2.9M	0
8	3,172	0	537,565	0	328,696	0	16,283	0	561,373	322	2.2M	0
9	1,530	0	412,480	0	248,682	0	8,899	0	448,728	25	1.8M	0
10	776	0	328,703	0	194,327	0	5,146	0	370,817	3	1.4M	0
11	424	0	265,642	0	155,877	0	3,097	0	310,242	0	1.2M	0
12	232	0	221,331	0	127,750	0	1,881	0	267,286	0	1.0M	0
13	133	0	187,072	0	105,924	0	1,120	0	232,668	0	859,799	0
14	87	0	159,780	0	89,628	0	786	0	199,201	0	739,345	0
15	49	0	138,268	0	77,040	0	470	0	178,118	0	640,093	0

Table 17: Frequencies of K-2-Paths in Empirical Versus Random Networks (MAG and MEDLINE)

k	**MAG**						**MEDLINE**					
	Algorithmic		**First-Initial**		**All-Initials**		**Algorithmic**		**First-Initial**		**All-Initials**	
	Empirical	Random	Empirical	Random	Empirical	Random	Empirical	Random	Empirical	Random	Empirical	Random
1	17.7M	39.2M	135.6M	72.6M	86.6M	57.0M	12.3M	28.2M	161.2M	45.6M	117.2M	35.7M
2	2.7M	5,387	10.7M	42,429	7.8M	18,107	2.0M	4,414	16.0M	21,014	10.2M	8,922
3	702,060	1	3.0M	18	2.1M	2	560,035	0	5.0M	3	2.9M	1
4	248,020	0	1.2M	0	863,951	0	211,491	0	2.2M	0	1.1M	0
5	99,790	0	634,831	0	427,499	0	85,655	0	1.2M	0	596,534	0
6	42,535	0	367,595	0	238,672	0	36,202	0	739,380	0	341,147	0
7	19,039	0	232,477	0	146,969	0	14,594	0	493,980	0	215,977	0
8	8,394	0	158,775	0	97,664	0	5,346	0	353,083	0	146,858	0
9	4,089	0	114,152	0	68,906	0	1,625	0	265,558	0	105,815	0
10	1,884	0	84,754	0	49,841	0	466	0	206,788	0	80,431	0
11	949	0	65,889	0	38,540	0	139	0	166,021	0	62,975	0
12	532	0	52,538	0	30,110	0	33	0	137,250	0	50,460	0
13	309	0	42,106	0	24,038	0	24	0	114,121	0	40,643	0
14	206	0	35,186	0	19,799	0	8	0	96,875	0	34,081	0
15	119	0	29,179	0	16,313	0	4	0	83,126	0	28,455	0

5.16 VULNERABILITY OF NETWORK MEASURES TO NAME AMBIGUITY

The main finding from analyzing the impact of author name disambiguation on network properties is that depending on the choice of disambiguation method we can understand the same data in different ways. In addition, the same data can be distorted to different levels depending on operationalization of measures. Table 18 summarizes the average error ratio per measure and thereby enables the ranking of each measure per initial based method in an ascending order of vulnerability. Here, vulnerability means the absolute value of average error ratios across four datasets per measure. For instance, the number of unique edges in a network disambiguated by all-initials shows the least vulnerability, while centralization by first-initial method is most vulnerable to name ambiguity. This implies that scholars who analyze networks compromised by name ambiguity can select less vulnerable measures to better approximate the ground-truth network properties and to take extra caution when their research involves one or two measures that are highly vulnerable to name ambiguity.

Table 18: Average Error Ratio (% or %p) of Measures

Metrics	Disambiguation Method	Data				Vulne-rability Rank
		DBLP	KISTI	MAG	MEDLINE	
No. of Unique Authors	First-Initial	-29.70	-94.01	-33.40	-38.67	13
	All-Initials	-17.81	-85.88	-20.97	-20.23	9
Average Production	First-Initial	44.81	626.37	53.53	65.44	20
	All-Initials	22.83	209.86	27.69	25.97	17
Production Gini	First-Initial	12.88	74.83	32.62	24.44	8
	All-Initials	6.56	47.30	19.62	12.03	6
No. of Unique Edges	First-Initial	-2.28	-47.50	-6.97	-1.18	2
	All-Initials	-0.95	-3.41	-5.39	1.90	1
Average Degree	First-Initial	40.20	278.08	41.22	62.72	19
	All-Initials	21.28	197.01	20.32	28.11	16
Centralization	First-Initial	699.47	10,453.65	291.10	707.06	24
	All-Initials	367.84	5,023.39	190.23	427.99	23
Density	First-Initial	104.79	2,769.46	118.78	171.58	22
	All-Initials	49.96	868.40	54.34	61.82	21
Ratio of Largest Component	First-Initial	16.45	12.73	34.94	11.32	5
	All-Initials	10.26	18.65	22.09	7.38	3
Average Shortest Paths	First-Initial	-31.50	-60.93	-39.53	-37.39	11
	All-Initials	-20.53	-56.99	-27.73	-25.25	7
Assortativity	First-Initial	-5.28	-319.13	-41.77	-29.55	18
	All-Initials	-1.85	-196.29	-12.93	23.74	12
Transitivity (1-mode)	First-Initial	-52.27	-41.65	-53.84	-71.80	15
	All-Initials	-36.89	-71.58	-39.43	-58.59	14
Transitivity (2-mode)	First-Initial	-29.89	45.19	-31.95	-52.92	4
	All-Initials	-24.31	-62.46	-20.33	-56.99	10

CHAPTER 6: SIMULATION OF MERGING AND SPLITTING

The previous chapters showed that initial-based author name disambiguation can distort coauthorship network properties, and such a distortion was not trivial when compared to algorithmically disambiguated networks, which were used as proxies of ground-truth. Most distorted properties were attributed to the effects of merging, although splitting must have also affected the distortion. The nature of such impact was, however, not clearly explained. This chapter attempts to address how much merging or splitting can produce what levels or magnitudes of distortive effects on network properties. For this, the merging and splitting levels were simulated by an increment of 1% (from 0 to 100%) and, for each simulation, network measures that have been widely used in bibliometric studies were calculated. During this process, what levels of merging or splitting can be acceptable considering the errors of network measures induced by name ambiguity are also considered. A problem here is that it can be hard to reach a consensus about the acceptable level of name ambiguity in practice. The acceptable level of ambiguity can vary depending on the purpose and situations of individual studies. In this study, name ambiguity (merging or splitting) resulting in a 5% error of any network measure is set arbitrarily as an acceptable level for the purpose of illustration.

With this 5% error of network measurement in mind, the task of this chapter is to find what level of merging or splitting produces such an error in each network measure. For this purpose, steps for simulating merging and splitting effects were methodologically adopted from Fegley and Torvik (2013); Wang, Shi, McFarland, and Leskovec (2012). (1) A list of unique authors from algorithmically disambiguated data was obtained. (2) Each unique author in the list was assigned a name ambiguity label for the misidentification rate, with one of “merging,” “splitting,” or “both merging and splitting” per initial-based disambiguation method (the labels are mutually exclusive and only one label gets assigned to an author identity). (3) A set of unique authors was randomly selected from the list to result in an *N* % (from 1% to 100%) selection of unique authors who have the target ambiguity type. (4) Name instances associated with the randomly selected unique authors were changed to the first-initial or all-initials format, which replaced corresponding unique author IDs in algorithmically disambiguated data. This setup results in a dataset where an *N* % of unique authors were merged and/or split per initial-based disambiguation. (5) A network measure was calculated both for the algorithmically disambiguated data and the same data compromised by author name ambiguity to find what level of errors was induced by the *N* % of merged and/or split authors. (6) The steps from (1) to (5) were repeated until the *N* % of merging or splitting resulting in 5% of measurement error was found.

The numbers of four M-rate types per disambiguation method for four datasets are summarized in Table 19 (Type A = no compromise, Type B = pure merging, Type C = pure splitting, Type D = merging & splitting). In DBLP, for example, a total of 817,628 unique authors were identified by algorithmic disambiguation. Among them, approximately 373,000 authors (46%) were not compromised by the first-initial disambiguation method, while 444,597 author identities (54%) were merged and/or split. For acceptable error detection, the *N* % of authors whose identities were compromised by merging (Type B and Type D) was randomly selected and changed into initial format for given names.

Table 19: Summary of M-Rate Type Frequencies Per Disambiguation Method

Data	Disambiguation Method	Number of Authors & Ratio	M-Rate Label				Total
			Type A	Type B	Type C	Type D	
DBLP	First-Initial	No. of Authors	373,031	444,597	0	0	817,628
		%	45.62	54.38	0	0	100
	All-Initials	No. of Authors	486,250	331,378	0	0	817,628
		%	59.47	40.53	0	0	100
KISTI	First-Initial	No. of Authors	23,667	252,599	506	33,108	309,880
		%	7.64	81.52	0.16	10.68	100
	All-Initials	No. of Authors	35,037	226,912	1,905	46,026	309,880
		%	11.31	73.23	0.61	14.85	100
MAG	First-Initial	No. of Authors	330,974	502,262	0	0	833,236
		%	39.72	60.28	0	0	100
	All-Initials	No. of Authors	460,160	373,076	0	0	833,236
		%	55.23	44.77	0	0	100
MEDLINE	First-Initial	No. of Authors	722,145	1,073,772	18,018	25,472	1,839,407
		%	39.26	58.38	0.98	1.38	100
	All-Initials	No. of Authors	922,523	752,651	82,458	81,775	1,839,407
		%	50.15	40.92	4.48	4.45	1

Regarding error simulation hereafter, two points are worth noting. For merging error simulation, Type B (pure merging) and Type D (merging & splitting) were used for random selection of authors. For splitting, however, only Type C (pure splitting) was used. This is because the majority of Type D cases consist of most cases merging with a very small number of splitting. Specifically, when ten name instances of a unique author in a dataset are both merged and split by initial-based disambiguation, nine instances are merged and one is split. This means that, if Type D is considered for splitting error simulation, the merging effect will be more pronounced than splitting, thus blurring the impact of splitting. Another note is that splitting error simulation for KIST was not conducted because the number of unique authors who were compromised by splitting was very small: 506 (0.16%) by the first-initial method and 1,905 (0.61%) by the all-initials method.

Figure 20-23 (four figures) show the outcomes of nine network measures when *N* % of merging happens randomly in four datasets per disambiguation method: first-initial method (green triangles) and all-initials based one (red crosses). An inset figure shows the levels (i.e., percentages) of merging that induces 5% or less error when compared to the outcomes calculated for networks disambiguated by algorithms. The maximum percentage of merging producing 5% or less error is recorded in green (first-initial) or red (all-initials).

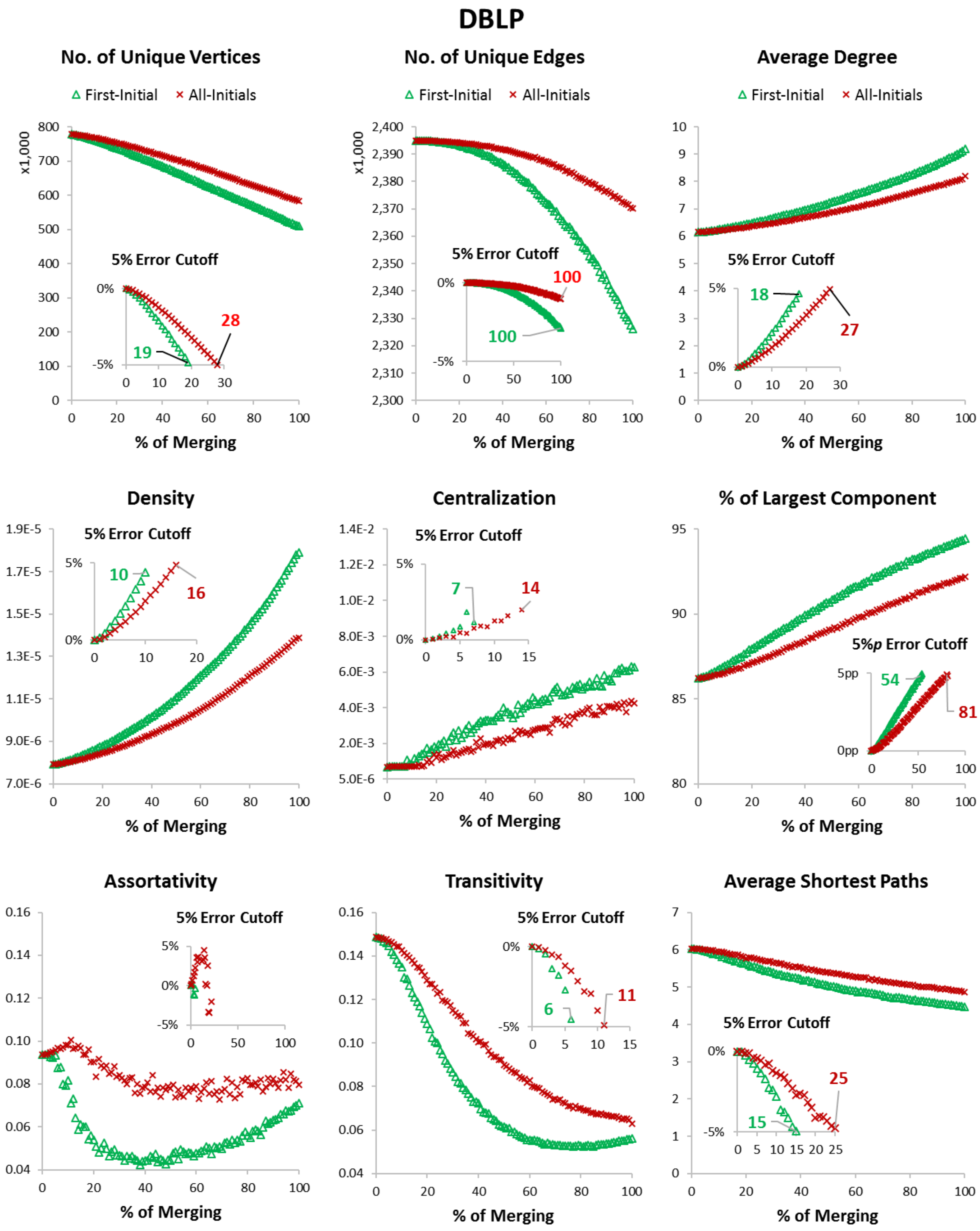


Figure 18: Change in Measures for DBLP Given Merging Level

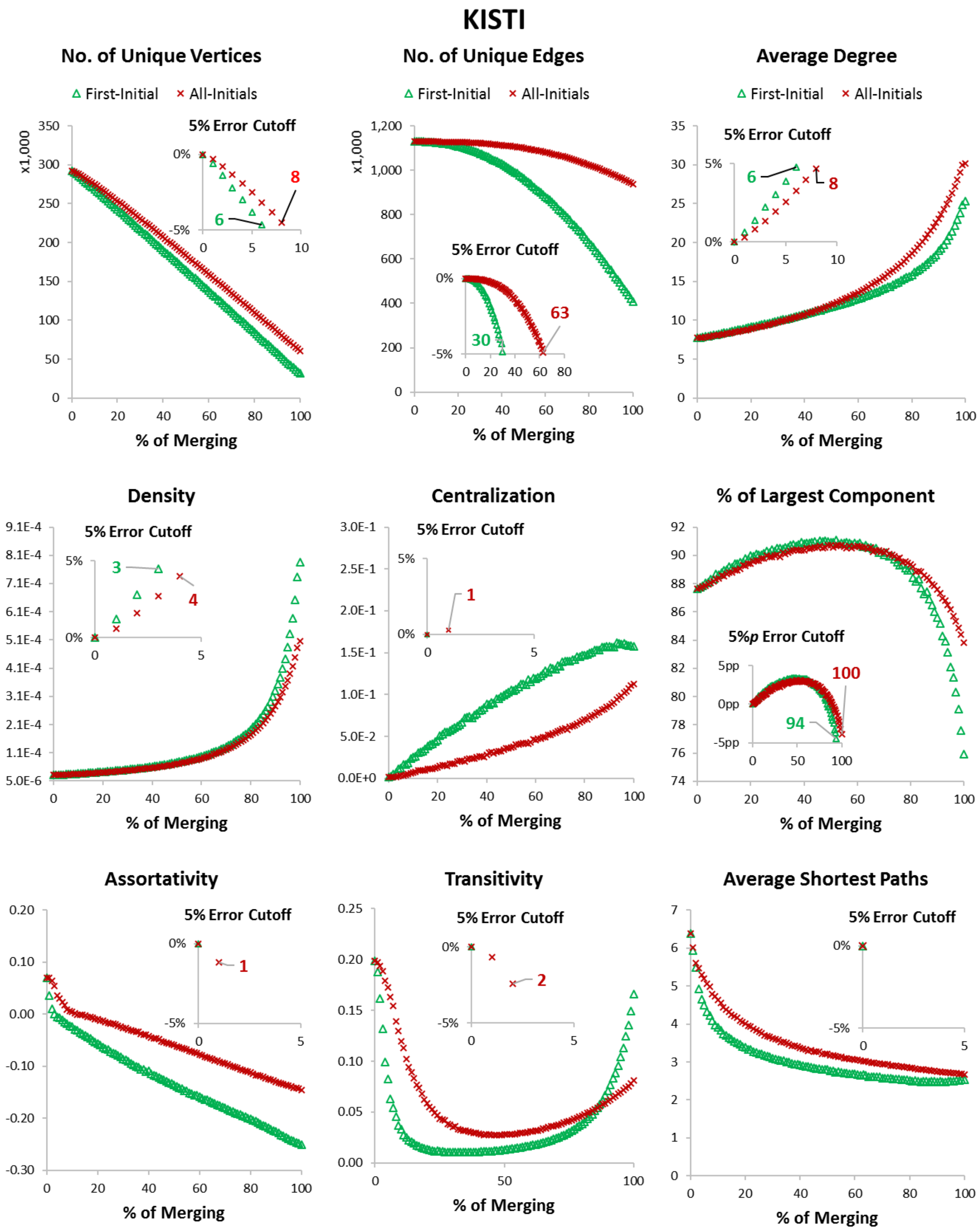


Figure 19: Change in Measures for KISTI Given Merging Level

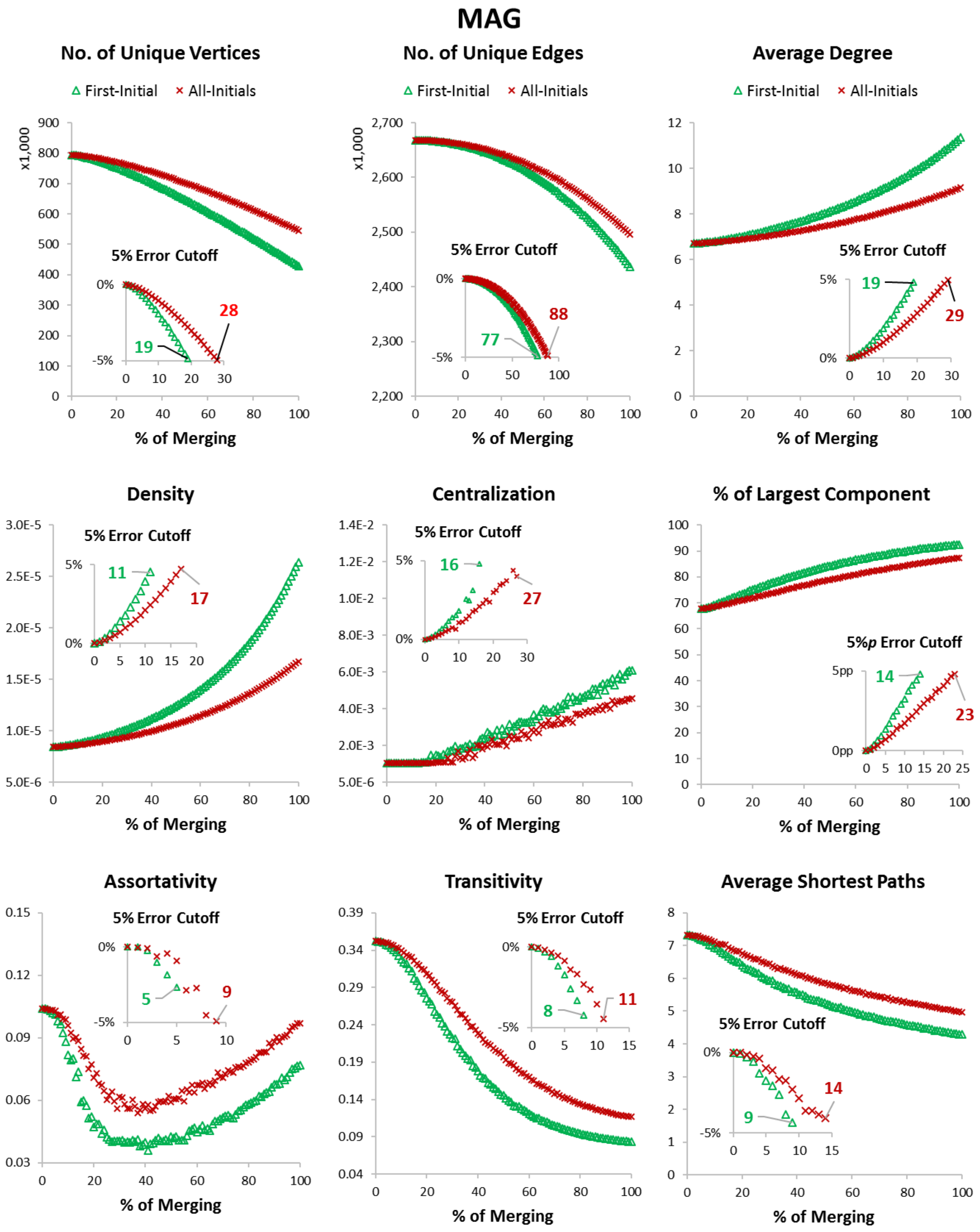


Figure 20: Change in Measures for MAG Given Merging Level

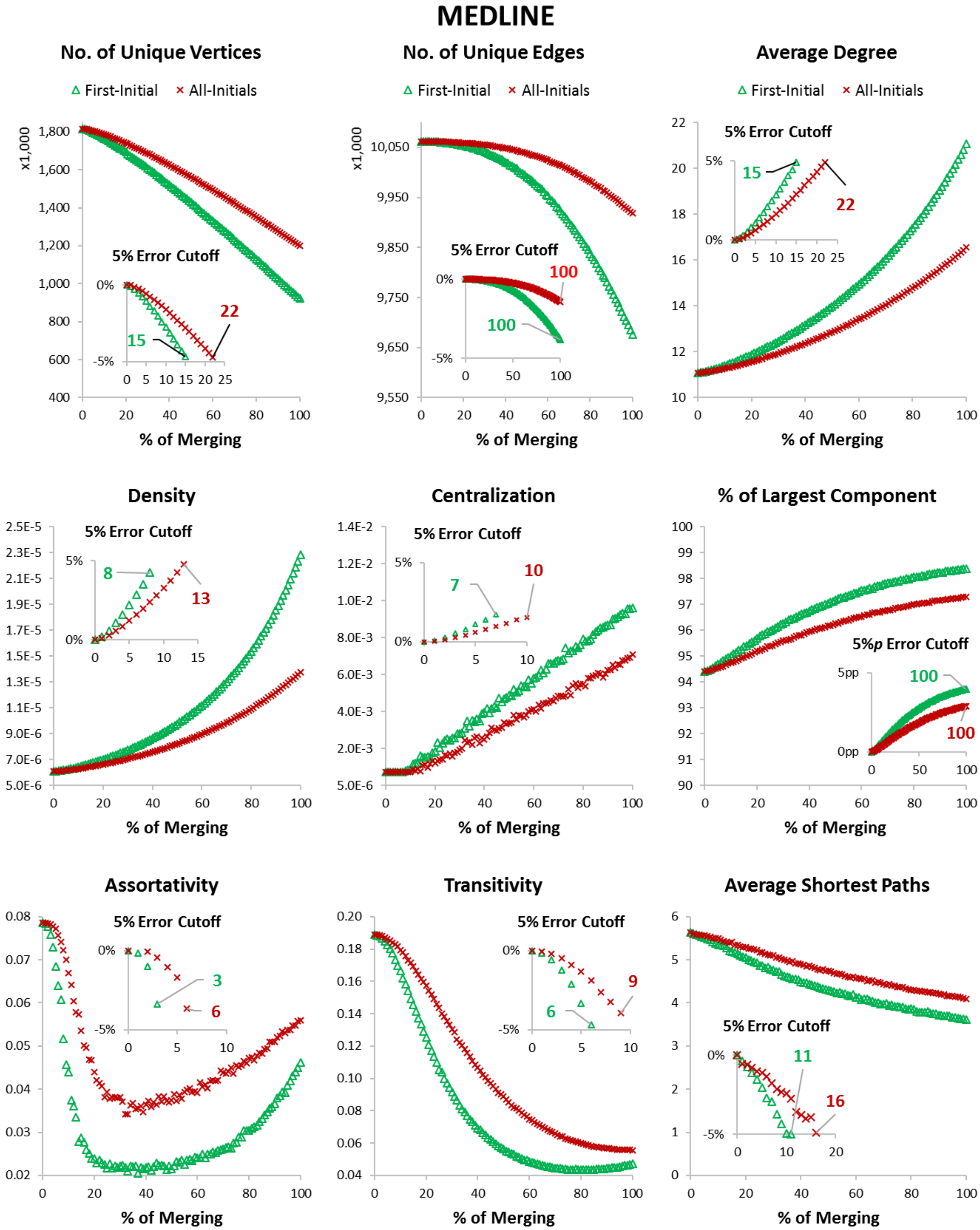


Figure 21: Change in Measures for MEDLINE Given Merging Level

All figures for merging simulation show similar trends. First, as the merging ratio increases from zero to 100%, the values of some measures increase (plots move upward and right). These measures are average vertex degree, network density, degree centralization, and the ratio of the largest component (except for KISTI). Second, other measures such as the number of unique authors, the number of edges, transitivity, and average shortest path lengths decrease as the merging ratio increases. These tendencies are consistent with findings from the previous chapter on the effects of initial-based disambiguation on network metrics, where merging was dominant when network datasets were pre-processed by the first-initial or the all-initials methods. Third, the change induced by the first-initial method is greater than that by the all-initials method across most network measures. This is because the first-initial method produces more merged authors than the all-initials method, as shown in Figure 3. This means that, given the same *N* % merging level, the first-initial method introduces a larger number of merged authors than the all-initials method, leading to larger impact on network properties.

Several exceptional cases are worth a further explanation. First, degree assortativity shows U-shape changes in DBLP, MAG, and MEDLINE. Degree assortativity in this study was calculated by using the Pearson's correlation coefficient as described in Newman (2001b). According to this method, assortativity increases when vertices having similar degrees tend to be linked to each other. This assortativity calculation is, however, known to be sensitive to vertices with high degrees. For instance, Fegley and Torvik (2013) suggested that, even if most vertices in a network have neighboring vertices of degree similar to them, the overall degree assortativity can be deflated by outliers with high degrees who have many neighbors with low degrees. Thus, the U-shape change of assortativity can be conjectured to form because degree similarity among high-degree vertices changes due to increased merging. Specifically, to some ratios of merging, vertices of small degree were merged to produce high-degree vertices with neighbors of low degree, which reduces the overall assortativity. Then, higher levels of merging began to connect vertices of high degree, leading to an increase in the assortativity level. Second, the assortativity in KISTI develops consistently toward a negative correlation. This means that merging leads to a situation where high-degree vertices tied to low-degree vertices are dominant and their structure of connection is like a star network (i.e., a vertex is connected to many vertices that are disconnected from each other). Third, the first-initial method showed different patterns of change for centralization, ratio of the largest component, and transitivity, when compared to DBLP, MAG, and MEDLINE. This can be because of the dramatic change in the number of vertices and edges due to exceptionally high levels of name ambiguity that is unique to Korean names as illustrated by Figure 4, Figure 7, and Case B in Table 10.

Regarding the acceptable error rate (5% cutoff), some measures were more tolerable to merging than others. For example, the number of edges showed less than 5% of errors by 100% of merging by initial-based method in DBLP (top-middle in Figure 18) and MEDLINE (top-middle in Figure 21). For MAG, 77% and 88% of merging produced 5% of error. Unlike the edge count, transitivity in DBLP (bottom-middle in Figure 18), MAG (bottom-middle in Figure 20), and MEDLINE (bottom-middle in Figure 21) allowed only 6~11% of merging by initial-based disambiguation for the 5% error rate. In KISTI, a very small percentage of merging could exceed the proposed acceptable error in most measures.

Unlike merging, the effects of splitting on network properties were more pronounced by the all-initials method than by the first-initial method. Figure 22 shows the change of network measures under various splitting ratios per initial-based disambiguation for MEDLINE. Here, changes in network measures by the all-initials method (red crosses) are larger than those by the first-initial method (blue triangles). This is because the all-initials method captures name variants more frequently than first-initial method, as shown in Figure 3: when disambiguated by all-initials, KISTI and MEDLINE showed higher level of splitting (Type C – pure splitting and Type D – merging and splitting). Since the number of unique authors vulnerable to splitting is larger by all-initials than by first-initial, the same *N* % splitting for simulation produced a larger number of splitting cases for all-initials method than for first-initial method, which leads to higher error rate due to splitting.

Also, unlike merging, splitting has a limited impact on network properties. A noticeable observation is that even with 100% of splitting, many measures led to less than 5% error compared to values obtained from the algorithmically disambiguated MEDLINE. For example, when disambiguated by first-initial method, 100% of splitting does not exceed 5% error in all of nine measures. By all-initials method, six measures – number of unique vertices, number of unique edges, average vertex degree, degree centralization, ratio of the largest component, and transitivity - show less than 5% error in approximating the proxy of ground truth by algorithmically disambiguated MEDLINE. This indicates that, when compared to errors produced by merging in Figure 21, network measurement errors due to initial-based disambiguation were less sensitive to splitting than to merging. This observation confirms the findings from Fegley and Torvik (2013) and Wang et al. (2012).

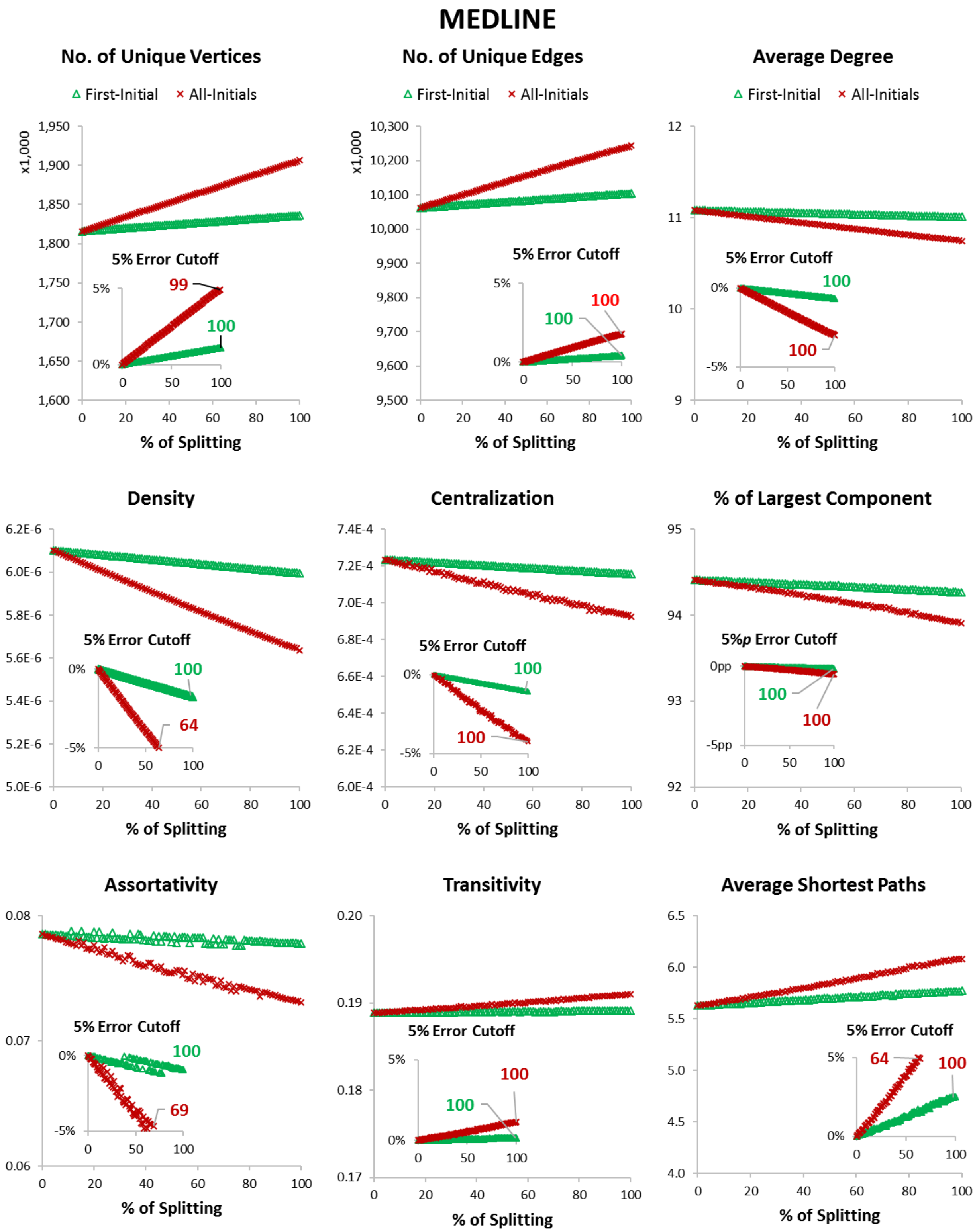


Figure 22: Change in Measures for MEDLINE Given Splitting Level

Most importantly, the changes of most network measures per initial-based disambiguation were almost linear. This means that, given a specific level of splitting, the magnitude of error induced by splitting could be estimated with high accuracy. This situation, in contrast, does not apply to merging. While the merging levels acceptable for a given error rate (5% here) could be found by simulation, the curvature of plots in the merging simulation above implies that the estimation of error rate based on a specific merging level is challenging. In other words, changes of network measures due to merging were not proportional to the merging ratios. As noted in Fegley and Torvik (2013), the network measures under varied merging levels tend to change in a non-linear way. In addition, even for the same measure, the change curves were different depending on datasets, implying that merging-induced changes of network measures may be dependent on the characteristics, e.g., topologies or domains, of individual networks and hard to generalize for application to other datasets.

CHAPTER 7: ESTIMATION OF AMBIGUITY LEVEL

Previous chapters showed that initial-based disambiguation of bibliometric data can lead to the distortion of network properties and sometimes to false positive findings from the same data. Through simulation, the severity of distortion was shown to vary depending on the disambiguation method (i.e., first-initial or all-initials), the type of name ambiguity (i.e., merging or splitting) and the level of each ambiguity type. Given an acceptable level of measurement errors induced by initial-based disambiguation, the ratio of merging or splitting could be found by referring to the change curves of network measures.

This chapter addresses the possibility of estimating the presence and magnitude of name ambiguity from scholarly data. Specifically, it explores whether the level of merging or splitting caused by initial-based disambiguation can be inferred from features extracted from the target data. Considering that scholars rely heavily on initial-based disambiguation for analyzing bibliometric data, the estimation of merging and splitting levels can help scholars gauge, even if roughly, what levels of measurement errors their analysis of a bibliometric dataset is possibly prone to.

The first step to the analysis was to generate 100 sets of randomly selected papers ranging from 1,000 to 100,000 with an increment of 1,000. For each set of papers, the number of unique authors who are merged or split by first-initial or all-initials method was counted (integer). This process was consistently used in this thesis research as the indicator of name ambiguity. Also, the ratio of the merged or split authors over the total number of unique authors was calculated (ranging from 0 to 1). These two values were used as the dependent variables.

Several features were extracted from each set of papers for independent variables. First, the number of selected papers was chosen (integer). I base this choice on the hypothesis that distortion of coauthorship network properties is positively associated with the size and comprehensiveness of a network (Fegley & Torvik, 2013). Next, the numbers of given name strings in full, one-initial, and two or more initials were counted respectively (three integers). This selection is based on the observation that author names with full given name string would reduce name ambiguity (Han et al., 2005; Torvik & Smalheiser, 2009).

A third type of feature was the ethnicity associated with author names. A unique author in a selected paper was matched with an ethnicity assigned by Ethnea, an ethnicity classifier (Torvik & Agarwal, 2016). Ethnea assigns a class of ethnicity (among 26 ethnicities) to a name instance by first searching the name in a database where an author name instance is assigned to a country based on the geo-code information of an affiliation associated with the name instance (Torvik, 2015) and then mapping the country distribution to ethnicities using a logistic regression model (for more details, refer to Torvik and Agarwal (2016)). Among 26 ethnicities, those appearing for more than 1% of unique authors in the whole dataset were counted in each set of sampled papers (13~15 integers). The inclusion of ethnicity is based on the observation that some names originating from specific regions, e.g., China and Korea, tend to share common surnames and given names, thus being more ambiguous than names from other regions (Kim & Diesner, 2016; Strotmann & Zhao, 2012; Torvik & Smalheiser, 2009).

Table 20 lists the types of features generated for each data. Regarding feature generation, two exceptions were made. First, in DBLP, unique authors are represented by name strings (name = ID). On average, 97% of author IDs have full given names. Second, in KISTI, about 98% of unique authors are Korean because the data contains publication records published by domestic journals and conferences. Thus, given name string features were not considered for DBLP and ethnicity was not included as a feature for KISTI.

Table 20: Summary of Features for Estimation

Feature	DBLP	KISTI	MAG	DBLP
Name Ambiguity	Merging	Merging Splitting	Merging	Merging Splitting
Given Name String	N/A	Full, One Initial, Two or More Initials	Full, One Initial, Two or More Initials	Full, One Initial, Two or More Initials
Ethnicity (In order of frequency)	Chinese, English, German, Hispanic, Indian, Japanese, French, Italian, Arab, Slav, Korean, Nordic, Greek, Dutch, Israeli	N/A	English, German, Japanese, Chinese, Hispanic, Italian, French, Nordic, Slav, Indian, Dutch, Korean, Arab	English, Japanese, German, Chinese, French, Hispanic, Italian, Slav, Nordic, Dutch, Indian, Korean, Arab

Exploratory analyses by a standard multiple regression for four datasets revealed that all variables are highly correlated with one another (Pearson's *r* = 0.97 and above). Such a high correlation indicates that feature values are quite evenly distributed across data. This also implies that one representative independent variable can be used for estimating the dependent variable. The number of papers showed the highest level of correlation with others consistently across disambiguation method and datasets. Thus, this metric was chosen as a simple model of predicting the number of merged or split authors. Figure 23 shows the result of curve fitting for the number of merged authors as a function of the number of papers per initial-based disambiguation method: green triangles for first-initial method and red crosses for all-initials method. The fitted lines – linear (i.e., bivariate regression) or polynomial – are shown along with equations and R-squared values per disambiguation method: a solid line for first-initial method and a dashes line for all-initials method.

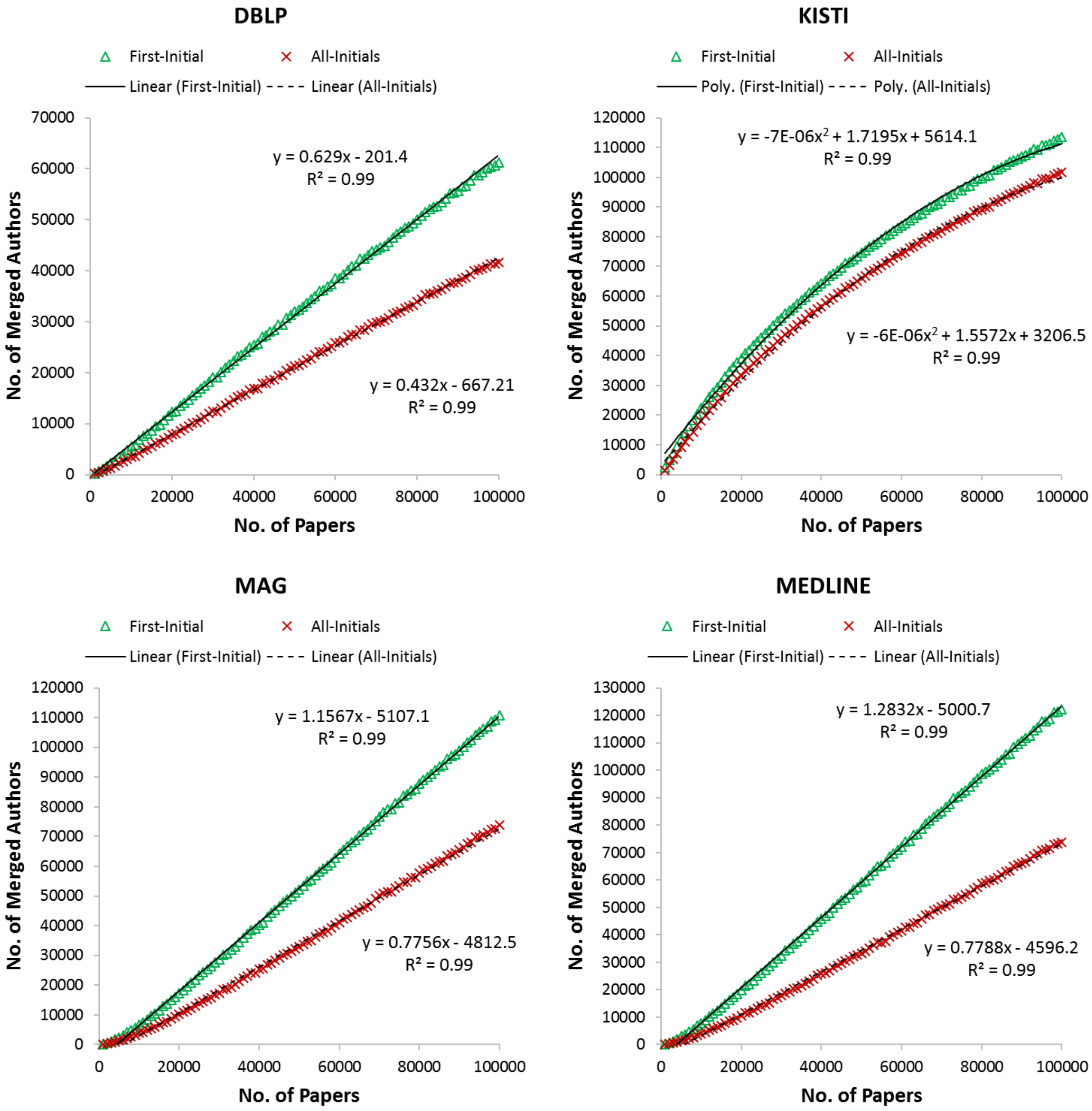


Figure 23: Curve Fitting for the Number of Merged Authors

The results show that the number of merged authors may be estimated quite accurately by the number of papers: R-squared values are all .99 and above, meaning that more than 99% of variance in the number of merged authors can be explained by the number of papers. This means that once we know the number of papers in a dataset, we can estimate how many author identities are merged quite accurately. For DBLP, MAG, and MEDLINE, a simple linear relationship by bivariate regression was the best fitting model, while for KISTI, the relationship between independent and dependent variables are best explained by a polynomial curve.

In Figure 24, curve fitting for the number of split authors as a function of the number of papers shows the same finding. In KISTI, the first-initial based disambiguation produced split authors in a linear way (R-squared = .96), while the all-initials in a polynomial curve (R-squared = .99). In MEDLINE, both initial-based methods took the simple linear form in producing split authors (both R-squared = .98).

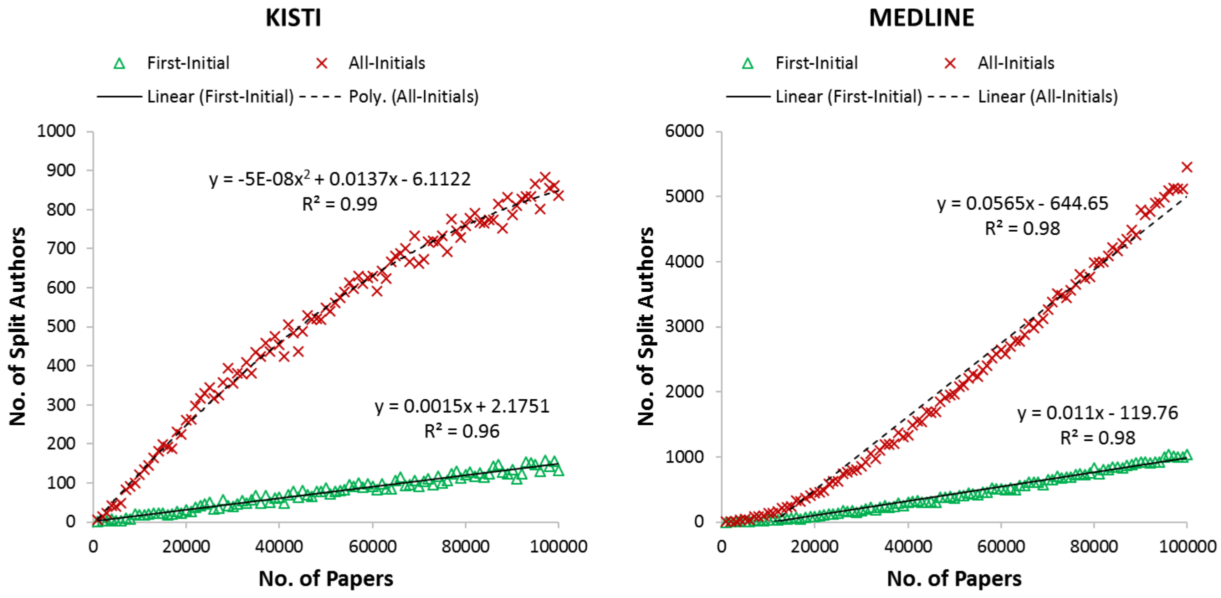


Figure 24: Curve Fitting for the Number of Split Authors

The findings above possibly confirm the hypothesis that the distortion of network measures is positively associated with the data size and comprehensiveness of a network (Fegley & Torvik, 2013). If this increase of merged or split authors (= numerator) proportional to the data size is greater than the increase of the number of unique authors (= denominator), the ratio of merged or split authors (= numerator/denominator) will also increase. As shown in previous chapter, the increased ratio of merged or split authors is related to more severe distortion of network measures. To check whether this distortion actually gets worse, the ratio of merged or split authors over the total number of unique authors per sample size was calculated. The results are plotted in Figure 25 for merging and in Figure 26 for splitting.

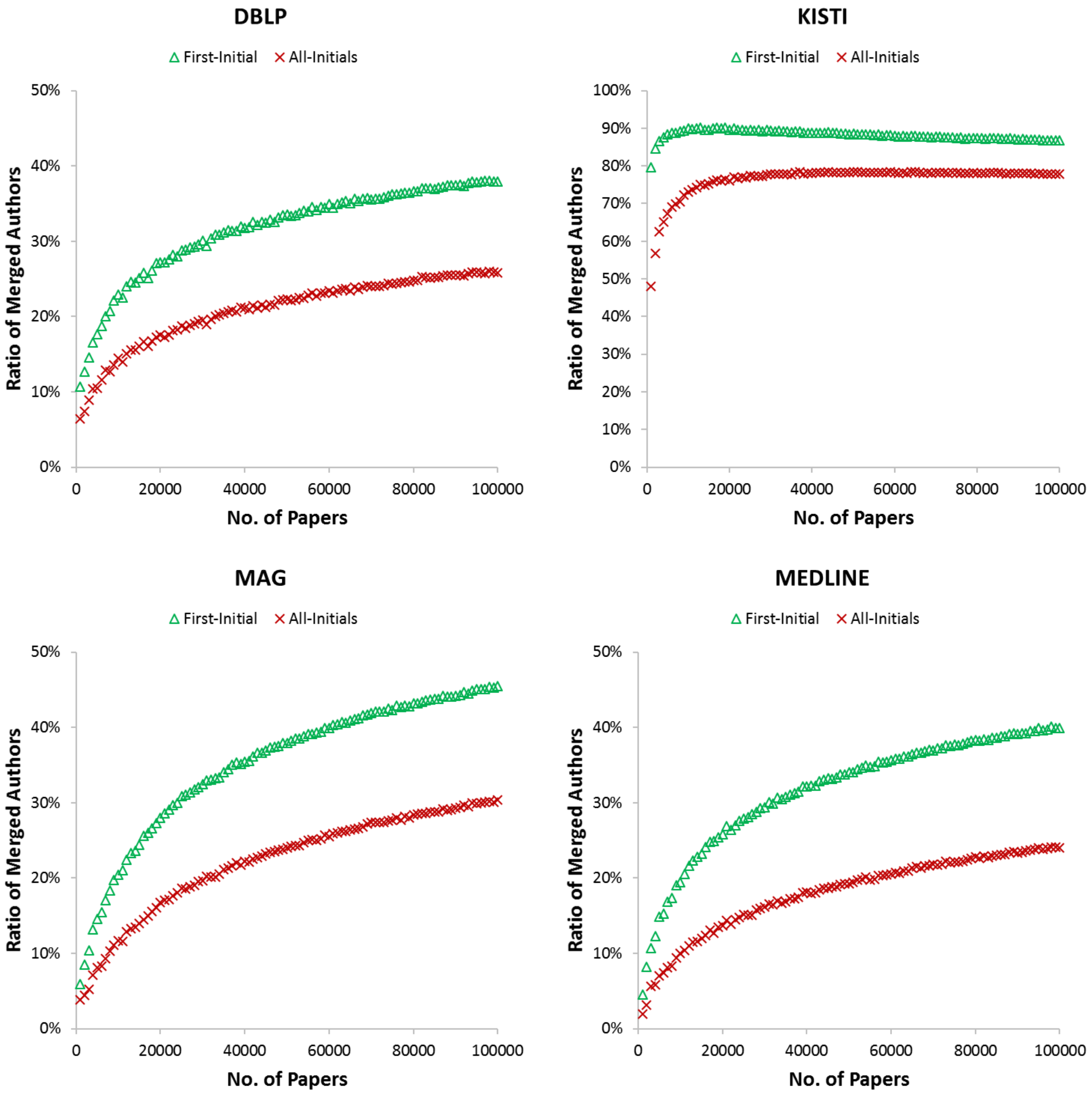


Figure 25: Ratio Change of Merged Authors Per Data Size

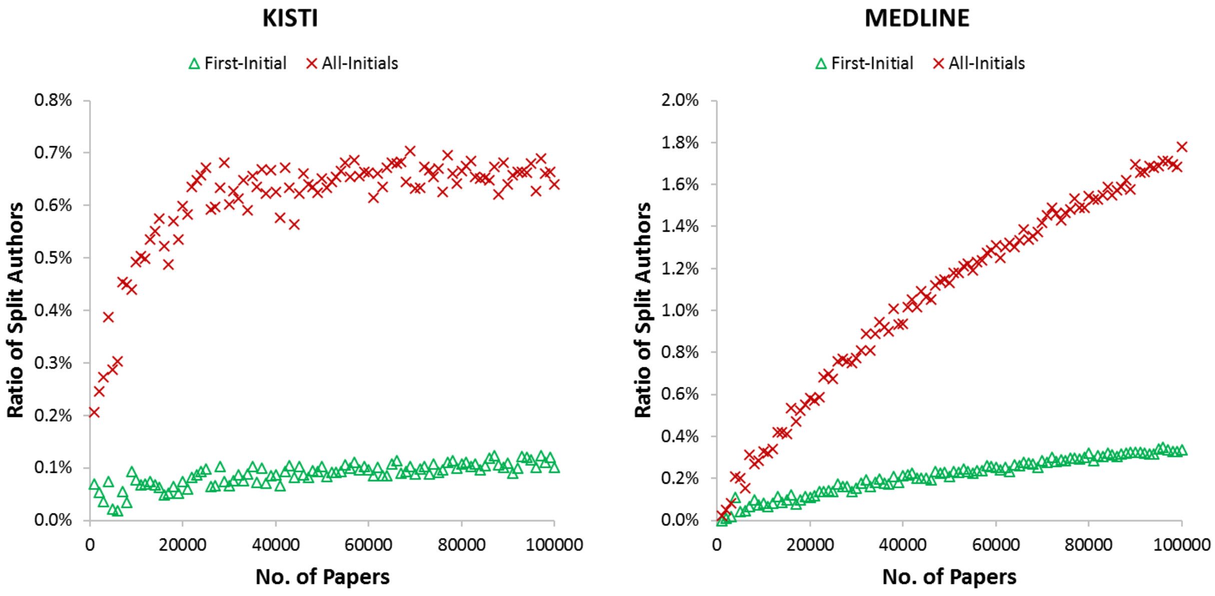


Figure 26: Ratio of Split Authors Per Data Size

The figures show that, as the data sizes increase, the ratios of merged or split authors over the total number of unique authors also increase. The changes of ratio were greater when the numbers of papers were smaller but, after the 20000 or 30000 paper size mark, tended to become plateaued, possibly leading to linearity. In KISTI, the ratios of merged authors for both initial-based methods hit ceilings (90% for the first-initial method and 75% for the all-initials method), and stabilized along or slightly below them, illustrating the characteristic of the KISTI data where most names become ambiguous when disambiguated by initial-based method.

CHAPTER 8: DETECTION OF MERGED AUTHORS

In the preceding chapter, the number of merged or split authors in a bibliometric dataset was estimated by the data size and the number of publications. This positive association between the number of merged/split authors and the data size leads to an additional question of whether merged or split authors can be identified so that corrective measures can focus on them. In other words, if merged or split authors can be identified, the name instances associated with the merged or split authors can be selectively disambiguated by computationally or manually, leading to a decrease in the name ambiguity level and, accordingly, the level of errors in network measurement.

This chapter tests this strategy using ego-network properties including degree, density, and local clustering coefficient. The selection of predictors were guided by the findings from previous chapters. First, the change trends of average degree in Figure 8 (Trend of Average Degree) and Figure 18 ~ Figure 21 (see subfigures for Average Degree in Change in Measures for four datasets per merging level) imply that nodes representing merged authors are likely to have a higher degree than other nodes representing non-merged authors, although some nodes representing authors who are not merged may also have a high degree because of frequent collaboration with diverse coauthors or being involved in a large-scale coauthorship. Second, as explained for the relatively low rate of change of unique edges for Figure 7 and Table 10, when multiple author nodes are merged into one, their coauthors are not likely to be merged together (with the exception of KISTI). In addition, we can hypothesize that those coauthors are not likely to collaborate with one another. This leads to the conjecture that a merged author would have low ego network density, which is the ratio of existing edges (almost constant) over possible edges (increasing as more coauthors are wrongly attached to the merged author) in an ego-network. Third, as shown in Figure 16 and Figure 18 ~ Figure 21 in previous chapters, transitivity showed variance in change as the number of merged authors increased. Since transitivity equals the average of individual authors' local clustering coefficients, a merged author is likely to have a low local clustering coefficient. This is also logically conjectured as merging and causes coauthors of multiple authors to be attached to a merged author (i.e., increased 2-paths) without forming edges among them (i.e., increased 2-paths that do not get closed).

A combination of these three independent variables (numeric) will be used to estimate the level of merging of authors disambiguated by initial-based method. Here, the dependent variable is the number of unique authors disambiguated algorithmically who are merged into an author identity disambiguated by either first-initial or all-initials method. For feature extraction, the data were disambiguated by an initial-based method. Second, a list of unique authors per initial-based method was created. Third, each author was assigned the number of unique authors who were in the original data but merged by initial-based disambiguation by referring to the authorship position and paper IDs associated with author instances in a) algorithmically disambiguated data as well as b) the same data, but disambiguated by given name initial(s). Next, ego-network metrics (degree, density, and local clustering coefficient) were calculated for each unique author identified by initial-based method in the list. Finally, a subset of 20,000 authors was selected from the list by randomly choosing 10,000 authors on the list who do not have a merged author identity at all and another 10,000 authors who have multiple merged identities.

Note that splitting was not considered for estimation. Detection of merged authors can be straightforward because a merged author identity by initial-based disambiguation contains feature information of merged identities. In contrast, to detect split authors, pairwise comparison of author identities together with feature values is required, which can bring in complexity for analysis in terms of operationalization of measurement. It was also considered that effects of splitting were found to be less distortive than merging.

Hierarchical multiple regression using two models was used to assess the ability of three independent variables –degree, ego-network density, and local clustering coefficient – to predict levels of merged author identities (i.e., how many author identities by algorithmic disambiguation are merged into a unique author identified by initial-based disambiguation). For the first model (Model 1), only degree was entered. The second model (Model 2) includes all the independent variables. Then, Model 1 and Model 2 were compared to measure the ability of ego-network density and local clustering coefficient to predict levels of merged author identities, after controlling for the influence of degree.

Table 21 reports regression results for four datasets. For example, in DBLP, the degree in Model 1 explains 84.1% of the variance ($\boldsymbol{R^2}$) in the levels of merged author identities by first-initial disambiguation method. After the addition of ego-network density and local clustering coefficient in Model 2, the total variance explained by the model was 84.3%. The two control variables explained an additional 0.2% of the variance ($\Delta\boldsymbol{R^2}$ for Model 2) in the levels of merged author identities, after controlling for degree. In the Model 2, all the variables were statistically significant (p <.000). Regarding Beta (Standardized coefficient), Degree recorded a higher value (.967) than Local Clustering Coefficient (-.049) and Density (.047). Standardized coefficients for all models in the table were statistically significant at p < .001.

Table 21: Model Summary of Hierarchical Regression for Merged Author Detection

Data	**Disambiguation Method**	**Model**	$\boldsymbol{R}$	$\boldsymbol{R^2}$	**Std. Error of Estimate**	$\boldsymbol{\Delta R^2}$	$\boldsymbol{\Delta F}$	**Sig.** $\boldsymbol{\Delta F}$	**Standardized Coefficients (β)**		
									Degree	**Density**	**Local CC**
DBLP	First-Initial	1	.917	.841	4.379	.841	105674.515	.000	.917	-	-
		2	.918	.843	4.347	.002	150.478	.000	.967	.047	-.049
	All-Initials	1	.951	.904	2.445	.904	188293.441	.000	.951	-	-
		2	.951	.905	2.428	.001	136.204	.000	.912	-.026	.045
KISTI	First-Initial	1	.813	.661	104.800	.661	35225.611	.000	.813	-	-
		2	.931	.867	65.670	.206	13952.115	.000	1.306	.033	-.659
	All-Initials	1	.941	.885	12.701	.885	154000.354	.000	.941	-	-
		2	.981	.963	7.222	.078	20930.229	.000	1.316	-.016	-.474
MAG	First-Initial	1	.909	.825	2.707	.825	94562.183	.000	.909	-	-
		2	.915	.838	2.609	.012	762.447	.000	.966	.023	-.122
	All-Initials	1	.913	.833	2.261	.833	99945.689	.000	.913	-	-
		2	.922	.849	2.148	.016	1077.397	.000	.974	.018	-.140
MEDLINE	First-Initial	1	.938	.879	4.553	.879	145147.111	.000	.938	-	-
		2	.942	.887	4.405	.008	680.921	.000	1.001	.081	-.057
	All-Initials	1	.923	.852	2.718	.852	115083.822	.000	.923	-	-
		2	.932	.870	2.551	.018	1348.098	.000	1.049	.053	-.163

The results suggest that three independent variables could quite accurately predict the levels of merged author identities by initial-based disambiguation: R-squared for Model 2 ranged from .838 (MAG, First-initial) to .963 (KISTI, All-Initials). In terms of Beta, Degree contributed most to the prediction followed by Local Clustering Coefficient. The direction of influence is negative for Local Clustering Coefficient with an exception for KISTI's All-Initials result. This is in line with a conjecture that, if an author by initial-based disambiguation contains many merged identities, the Local Clustering Coefficient would decrease. Most noticeable is, however, that Degree alone performed better than the combination of Density and Local Clustering Coefficient. The additional contribution by Density and Local Clustering Coefficient ($\Delta \boldsymbol{R}^{2}$ for Model 2) was between 0.1% and 20.6%. This implies that Degree can be used as a predictor of merged identities. These findings, however, should be accepted with care. Other network properties that were not tested in this thesis might perform better in predicting merged identities than the three variables. Also, how the test order of variables (e.g., Local Clustering Coefficient >> Density >> Degree) could affect the results was not tested.

CHAPTER 9: CONCLUSION AND DISCUSSION

9.1 SUMMARY OF ANALYSIS

This thesis first illustrates how certain choices for author name disambiguation in large-scale scholarly data can have an effect on our understanding of the properties of coauthorship networks and our reasoning about the mechanisms of coauthorship network evolution and coauthoring relationship formation. To investigate the importance of these effects, four large-scale scholarly datasets – DBLP, KISTI, MAG, and MEDLINE -- were obtained. Author names in the datasets "had been algorithmically disambiguated in a highly accurate fashion" or with decent accuracy (Kim & Diesner, 2015). Two initial-based methods for disambiguating author names – first-initial and all-initials – were applied to these datasets. These two methods were chosen because they have been widely used for resolving author name ambiguity in bibliometrics. Coauthorship networks were generated for each of algorithmic, first-initial, and all-initials name disambiguation approaches. Commonly used network metrics were calculated for each network, and their over-time changes were identified. In addition, the network properties were simulated with varying levels of merging and splitting and, for each level, network measures were calculated and compared.

Values of some network metrics showed an overall decrease when applying initial-based name disambiguation to proxies of ground truth datasets: the number of unique authors, assortativity, transitivity (both one-mode and two-mode metrics), and average shortest path lengths. This findings indicate that if researchers disambiguate author names using name initials, they "are likely to find coauthorship networks that are smaller, where people are closer to each other, less collaborative with shared coauthors, and less homogeneous in terms of collaboration partners than they actually are" (Kim & Diesner, 2016). Other measures' values increased: average production, average vertex degree, network density, and the ratio of the largest component. This implies that when using initial-based disambiguation, "scholars will appear to be more productive, collaborative, and imbedded in larger and more cohesive communities than they actually are" (Kim & Diesner, 2016). Two concentration measures – the Gini coefficient of author production and degree centralization – showed increases due to applying the initial-based method for data pre-processing. This suggests that, through the lens of initial-based disambiguation, resources and opportunities for research are more unevenly distributed in scientific communities than they actually are, and such inequality keeps increasing at a higher rate over time than what is accurate.

The general findings about these distortive effects of initial-based disambiguation based on this thesis are mostly in line with findings from previous research (Fegley & Torvik, 2013; Kim & Diesner, 2015, 2016; Wang et al., 2012). The added contribution with this thesis is that the impact of name disambiguation on networks and conclusions from network analysis was studied based on four large-scale datasets that differ in coverage of domains and disambiguation accuracy, and several common trends were found across datasets through over-time measurement and merging/splitting simulation. In addition, the levels of merging and splitting as a function of data size were estimated. Also, the ability of ego-network measures to predict the extent to which an author identity disambiguated by initial-based method may merge unique authors identified by algorithmic disambiguation was tested.

The main conclusion from this thesis is that "initial-based disambiguation can misidentify author identities mainly through merging, and, therefore, can distort macroscopic views of authorship patterns and the collaboration structure of a field or scientific community" (Kim & Diesner, 2016; Kim, Kim, & Diesner, 2014). In some cases, as shown for degree distribution, scholars who rely on initial-based disambiguation may arrive at false findings about the network topology (e.g., power-law distribution) and its generation mechanism (e.g., preferential attachment). As shown in the analysis of trend of transitivity and frequencies of *k*-2-paths, the choice for a disambiguation method can lead to different hypotheses about edge formation in coauthorship networks. Furthermore, "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" (Kim & Diesner, 2015).

9.2 PRACTICAL IMPLICATIONS

This study can serve as a warning signal to scholars and practitioners that they should be attentive to their choices of name disambiguation method when analyzing, curating, or reusing bibliometric data. Noticeably, several scholars who proposed and supported the usability of initial-based disambiguation, "began to perform algorithmic name disambiguation before conducting network analysis of large-scale bibliometric data (e.g., Deville et al., 2014; Martin et al., 2013)" (Kim & Diesner, 2016). The dominant practice in academia is, however, still initial-based disambiguation (Milojević, 2013; Strotmann & Zhao, 2012).

The dominance of the initial-based method is partially due to established practices by a few popular bibliometric data services. These services have been the main providers of bibliometric data for research through institutional subscription for decades. These services have provided data with author given names recorded in all-initials format in many cases. Findings from this study can be used to increase awareness of scholars and practitioners who have used scholarly data disambiguated by initial-based method and to help them improve disambiguation methods.

Another implication relates to the replication of discovered knowledge from bibliometric data. For decades, several bibliometric data services have been publicly available to scholars. This has created a situation where scholars can obtain a target dataset from multiple data sources. For example, a scholar who wants to study collaboration patterns among information scientists can download the metadata of publications published in a specific journal (e.g., *Journal of the Association for Information Science and Technology*) from ArnetMiner, DBLP, IEEE Xplore Library, Microsoft Academic Graph, SCOPUS, and Web of Science, to name a few. Each data service provides unique IDs or name strings to represent authors. Due to the different name disambiguation methods employed by the data services, the same journal papers can provide different, sometimes even conflicting, findings depending on the choice of data sources by a researcher. This situation can be called “Bibliometric Data Isomorphism,” a term (proposed by the author of this thesis) referring to a phenomenon where multiple versions exist for the same data points in bibliometric data. This problem also calls for the special attention of scholars and practitioners to bibliometric data quality in terms of author identification. For example, scholars may be encouraged to use computationally disambiguated data from, for example, ArnetMiner and DBLP for analyzing publication records in computer and information science.

9.3 SUGGESTIONS OF CORRECTIVE MEASURES

For a study using scholarly data, the preliminary step for scholars to take before analysis would be to resolve name ambiguity first, preferably by using high-performance algorithmic disambiguation that has been shown to approximate ground-truth data better than initial-based disambiguation (Kim & Diesner, 2015). Algorithmic name disambiguation, however, requires sophisticated knowledge of computation, proper feature selection, and careful implementation of algorithms, which may not be viable for scholars who have no adequate resources and capacities for such computational tasks. Another problem is that many tested algorithms were modeled to fit specific datasets (which can lead to overfitting); limiting their applicability to other datasets.

Based on the findings from this thesis, some corrective steps may be considered. First, scholars in need of analyzing bibliometric data disambiguated by initial-based method can refer to the relationship between the data size and the estimated level (ratio) of merged or split authors as in Figure 25. Given the level of merged or split author identities in a dataset, the error levels of network measures may be estimated by referring to distortion per merging or splitting level as described in Figure 18 ~ Figure 22. Although each dataset has domain-specific features (such as the average coauthor size or collaboration patterns), the reference to findings in this thesis can serve as a rough estimation of measurement error levels to help scholars decide whether to aim for disambiguation before analysis.

Another finding in this study helpful to scholars is that the levels of merged author identities could be predicted quite well by the degree, density, and local clustering coefficient of an author's ego-network. Scholars can apply these approaches for "focused disambiguation." Here is a suggestion. First, scholars can generate a list of authors identified from bibliometric data disambiguated by initial-based method. After networks are generated from the data and ego-network measures are calculated for each author on the list, they can apply the regression equation in previous chapter for detecting author nodes likely to contain merged identities. In a descending order of the levels of merged author identities, the author list is sorted. Depending on time and resources, the top *N* authors from the sorted list are selected, and their associated information such as full name string, coauthors, emails, or affiliation is used to disambiguate them manually or semi-automatically (e.g., if two author names share two or more coauthor names, they are presumed to relate to the same author identity). Once the disambiguation is done,

one or two network measures vulnerable to name ambiguity can be calculated for this partially disambiguated data and compared for difference to the values from the original data. These steps can be repeated until the change in differences of measures between *n*-times disambiguated data and the original data become small. As a heuristic, instead, authors with high degree might be selected without referring to any regression equation. This focused disambiguation can be effective if the level of merging or splitting in data is relatively low, but research using the data relies on network measures that are highly vulnerable to name ambiguity.

Another practical result from this thesis is the suggestion that splitting imposes less distortive effects on network properties than merging, which was also confirmed in Fegley and Torvik (2013) and Wang et al. (2012). This implies that, when disambiguating author names, strict rules for deciding matched name pairs are preferable to relaxed ones. When two name instances do not have much information for matching identities, it would be safer to regard them as separate identities than as the same one. This might explain why MAG allowed a higher level of splitting in disambiguating names as described in Table 3.

In academia, the collaboration between computer and information scientists and bibliometric scholars might be encouraged for ensuring proper control of name ambiguity in bibliometric data. First, this collaboration can be conducted by developing new algorithms or applying tested algorithms for disambiguating names to new research datasets. Second, scholars who have developed and tested methods for name disambiguation can deposit related code and disambiguated data (preferably with a report on disambiguation accuracy) in repositories for sharing them with other scholars. Currently, several bibliometric datasets are shared in data repositories such as Stanford Large Network Data Collection (https://snap.stanford.edu/data/), but they are not disambiguated at all or numeric IDs are assigned to authors after disambiguation by name initials. For sharing disambiguated data, especially, discussions on relaxed regulations may be required between scholars and bibliometric data service providers, which requires a further study by experts.

9.4 LIMITATIONS AND FUTURE DIRECTIONS

This study simulated the various levels of merging and splitting against four different datasets; a finding that disambiguation error can be estimated by the data size, i.e., the number of publications. However, little is known about how the data size of networks can affect the robustness of network measures to disambiguation error (merging and splitting) scenarios. In other words, given a data size (e.g., 100~100,000 randomly selected papers), vulnerability of network measures needs to be tested with varying levels of merging and splitting. This requires simultaneous manipulation of both the size of data and the level of ambiguity.

The impact of merging and splitting was investigated separately in this study. Since the impact of merging was more prominent than splitting across datasets, most of the distortion of network properties was explained in terms of merging. To obtain a better knowledge of the impact of name ambiguity, however, the interplay between the two disambiguation errors needs to be studied. For this, a future study needs to control both levels of merging and splitting and compare the changes of network measures accordingly.

In this study, the ethic or cultural background of authors (e.g., Chinese or Hispanic) was not considered much for estimating disambiguation errors. This was mainly because sets of randomly selected papers with different numbers provided similar distribution of name ethnicity per selection (refer to the high correlation among name ethnicity frequencies in Chapter 7). This does not mean that name ethnicity has no association with name ambiguity level. Scholars have suggested that specific ethnicities such as Chinese and Korean names may contribute to higher name ambiguity due to the naming culture of sharing common names (Kim & Diesner, 2016; Strotmann & Zhao, 2012; Torvik & Smalheiser, 2009). As shown for the case of KISTI, the impact of name ambiguity was severe across most network measures because of the significant amount of merging (i.e., more than 73% of all authors were merged in the whole KISTI data). This implies that, when such highly ambiguous names are more frequent than others, the distortion of network properties might be worse for the same size of data. This calls for a study where each name ethnicity is controlled for the frequency and the level of disambiguation errors to be tested for its impact on network properties. For example, a sensitivity test would need to be conducted to find out how much impact a certain type of name ethnicity can impose on a network if all names associated with the target ethnicity were correctly disambiguated, while names of another ethnicity remained ambiguous. The findings from this study are expected to provide practical insights such as ethnicity-focused disambiguation (e.g., first disambiguating all Chinese or Korean names).

Two-mode network approaches to studying coauthorship networks are still rare despite their theoretical justification (Newman et al., 2001; Opsahl, 2013). In this study, only transitivity was calculated on two-mode version of empirical and random networks. As shown in the Trend of Transitivity for Two-Mode Network (Figure 17), the two-mode approach revealed a subtle difference in understanding the impact of name disambiguation methods: while the disambiguated data showed a stable trend over time across four datasets, data disambiguated by name initials showed fluctuating trends. This finding was contrasted to the finding from one-mode network transitivity (Figure 16), where all disambiguation methods led to similar decreasing trends. This implies that, if tested by other two-mode network measures, disambiguation methods might show different impacts compared to when we use the one-mode network approach.

As such, this study can be viewed as a stepping stone for further studies on the impact of author name ambiguity on coauthorship network properties and is expected to help scholars establish better practices for knowledge discovery from ambiguous scholarly big data.

REFERENCES

Acedo, F. J., Barroso, C., Casanueva, C., & Galan, J. L. (2006). Co-authorship in management and organizational studies: An empirical and network analysis. *Journal of Management Studies, 43*(5), 957-983. doi:10.1111/j.1467-6486.2006.00625.x

Alstott, J., Bullmore, E., & Plenz, D. (2014). powerlaw: a Python package for analysis of heavy-tailed distributions.

Barabási, A. L., & Frangos, J. (2014). *Linked: the new science of networks science of networks*: Basic Books.

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. doi:10.1016/s0378-4371(02)00736-7

Bettencourt, L. M. A., Kaiser, D. I., & Kaur, J. (2009). Scientific discovery and topological transitions in collaboration networks. *Journal of Informetrics, 3*(3), 210-221. doi:10.1016/j.joi.2009.03.001

Bhattacharya, I., & Getoor, L. (2005). A latent dirichlet model for unsupervised entity resolution.

Biryukov, M., & Dong, C. L. (2010). Analysis of computer science communities based on DBLP. In M. Lalmas, J. Jose, A. Rauber, F. Sebastiani, & I. Frommholz (Eds.), *Research and Advanced Technology for Digital Libraries* (Vol. 6273, pp. 228-235). Berlin: Springer-Verlag Berlin.

Borgatti, S. P., Carley, K. M., & Krackhardt, D. (2006). On the robustness of centrality measures under conditions of imperfect data. *Social Networks, 28*(2), 124-136.

Börner, K., Maru, J. T., & Goldstone, R. L. (2004). The simultaneous evolution of author and paper networks. *Proceedings of the National Academy of Sciences of the United States of America, 101*(suppl. 1), 5266-5273. doi:10.1073/pnas.0307625100

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

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

Chua, A. Y. K., & Yang, C. C. (2008). The shift towards multi-disciplinarity in information science. *Journal of the American Society for Information Science and Technology, 59*(13), 2156-2170. doi:10.1002/asi.20929

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. doi:10.1002/asi.21363

Csardi, G., & Nepusz, T. (2006). The igraph software package for complex network research. *InterJournal*, 1695.

Culotta, A., & McCallum, A. (2005). *Joint deduplication of multiple record types in relational data.* Paper presented at the Proceedings of the 14th ACM international conference on Information and knowledge management.

De Nooy, W., Mrvar, A., & Batagelj, V. (2011). *Exploratory social network analysis with Pajek* (Vol. 27). New York: NY: 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*, 1-7. doi:10.1038/srep04770

Diesner, J. (2012), *Uncovering and managing the impact of methodological choices for the computational construction of socio-technical networks from texts*. Dissertations. 194. http://repository.cmu.edu/dissertations/194 .

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.

Diesner, J., Evans, C., & Kim, J. (2015). *Impact of entity disambiguation errors on social network properties*. Paper presented at the International AAAI Conference on Web and Social Media (ICWSM), Oxford, UK.

Ding, Y. (2011). Scientific collaboration and endorsement: Network analysis of coauthorship and citation networks. *Journal of Informetrics, 5*(1), 187-203. doi:DOI 10.1016/j.joi.2010.10.008

Fegley, B. D., & Torvik, V. I. (2013). Has large-scale named-entity network analysis been resting on a flawed assumption? *Plos One, 8*(7), 1-16. doi: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. 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

Frantz, T. L., Cataldo, M., & Carley, K. M. (2009). Robustness of centrality measures under uncertainty: Examining the role of network topology. *Computational and Mathematical Organization Theory, 15*(4), 303-328.

Glasser, G. J. (1962). Variance formulas for the mean difference and coefficient of concentration. *Journal of the American Statistical Association, 57*(299), 648-654. doi:10.2307/2282402

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

Guns, R., & Rousseau, R. (2014). Recommending research collaborations using link prediction and random forest classifiers. *Scientometrics, 101*(2), 1461-1473. doi:10.1007/s11192-013-1228-9

Han, H., Zha, H., & Giles, C. L. (2005). *Name disambiguation in author citations using a k-way spectral clustering method.* Paper presented at the Digital Libraries, 2005. JCDL'05. Proceedings of the 5th ACM/IEEE-CS Joint Conference on.

Holland, P. W., & Leinhardt, S. (1970). Method for detecting structure in sociometric data. *American Journal of Sociology, 76*(3), 492-&. doi:10.1086/224954

Illenberger, J., & Floetteroed, G. (2012). Estimating network properties from snowball sampled data. *Social Networks, 34*(4), 701-711.

Kim, J., & Diesner, J. (2015). The effect of data pre-processing on understanding the evolution of collaboration networks. *Journal of Informetrics, 9*(1), 226-236. doi:10.1016/j.joi.2015.01.002

Kim, J., & Diesner, J. (2016). Distortive effects of initial-based name disambiguation on measurements of large-scale coauthorship networks. *Journal of the Association for Information Science and Technology, 67*(6), 1446-1461. doi:10.1002/asi.23489

Kim, J., & Diesner, J. (2017). Over-time measurement of triadic closure in coauthorship networks. *Social Network Analysis and Mining*. doi:10.1007/s13278-017-0428-3

Kim, J., Diesner, J., Kim, H., Aleyasen, A., & Kim, H.-M. (Oct. 2014). *Why name ambiguity resolution matters for scholarly big data research.* Paper presented at the Big Data (Big Data), 2014 IEEE International Conference on.

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. doi:10.1633/JISTaP.2014.2.2.1

Kim, J., Tao, L., Lee, S.-H., & Diesner, J. (2016). Evolution and structure of scientific co-publishing network in Korea between 1948–2011. *Scientometrics, 107*(1), 27-41. doi:10.1007/s11192-016-1878-5

Larivière, 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

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

Martin, T., Ball, B., Karrer, B., & Newman, M. E. J. (2013). Coauthorship and citation patterns in the Physical Review. *Physical Review E, 88*(1), 012814-012811~012819. doi:10.1103/PhysRevE.88.012814

Merton, R. K. (1968). Matthew effect in science. *Science, 159*(3810), 56-&. doi:10.1126/science.159.3810.56

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:10.1016/j.joi.2013.06.006

Newman, M. E. J. (2001a). Clustering and preferential attachment in growing networks. *Physical Review E, 64*(2). doi:10.1103/PhysRevE.64.025102

Newman, M. E. J. (2001b). 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), 208701-208701~208704.

Newman, M. E. J. (2004). Coauthorship networks and patterns of scientific collaboration. *Proceedings of the National Academy of Sciences of the United States of America, 101*(suppl. 1), 5200-5205. doi:10.1073/pnas.0307545100

Newman, M. E. J. (2010). *Networks: An introduction*: Oxford University Press.

Newman, M. E. J., Strogatz, S. H., & Watts, D. J. (2001). Random graphs with arbitrary degree distributions and their applications. *Physical Review E, 64*(2), 026118.

Opsahl, T. (2009). *Structure and Evolution of Weighted Networks*. London, UK: University of London (Queen Mary College).

Opsahl, T. (2013). Triadic closure in two-mode networks: Redefining the global and local clustering coefficients. *Social Networks, 35*(2), 159-167.

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*. Paper presented at the Proceedings of the 9th ACM/IEEE-CS joint conference on Digital libraries, Austin, TX, USA.

Potamias, M., Bonchi, F., Castillo, C., & Gionis, A. (2009). *Fast shortest path distance estimation in large networks.* Paper presented at the Proceedings of the 18th ACM conference on Information and knowledge management.

Radicchi, F., Fortunato, S., Markines, B., & Vespignani, A. (2009). Diffusion of scientific credits and the ranking of scientists. *Physical Review E, 80*(5), 056103-056101~056110. doi:10.1103/PhysRevE.80.056103

Reitz, F., & Hoffmann, O. (2010). *Learning from the past: An analysis of person name corrections in DBLP collection and social network properties of affected entities.* Paper presented at the Advances in Social Networks Analysis and Mining (ASONAM), 2010 International Conference on.

Robins, G., Pattison, P., Kalish, Y., & Lusher, D. (2007). An introduction to exponential random graph (p*) models for social networks. *Social Networks, 29*(2), 173-191. doi:10.1016/j.socnet.2006.08.002

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

Sarawagi, S. (2008). Information extraction. *Foundations and Trends in Databases, 1*(3), 261-377. doi:10.1561/1900000003

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. doi:10.1007/s11192-014-1289-4

Shumate, M., & Palazzolo, E. T. (2010). Exponential random graph (p*) models as a method for social network analysis in communication research. *Communication Methods and Measures, 4*(4), 341-371. doi:10.1080/19312458.2010.527869

Sinha, A., Shen, Z., Song, Y., Ma, H., Eide, D., Hsu, B.-J., & Wang, K. (2015). *An overview of Microsoft Academic Service (MAS) and applications*. Paper presented at the Proceedings of the 24th International Conference on World Wide Web, Florence, Italy.

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. doi:Doi 10.1002/Asi.22695

Strotmann, A., Zhao, D., & Bubela, T. (2009). Author name disambiguation for collaboration network analysis and visualization. *Proceedings of the American Society for Information Science and Technology, 46*(1), 1-20.

Stumpf, M. P., & Porter, M. A. (2012). Critical truths about power laws. *Science, 335*(6069), 665-666.

Torvik, V. I. (2015). MapAffil: A bibliographic tool for mappign author affiliation strings to cities and their geocodes worldwide. *D-Lib Magazine, 21*. doi:10.1045/november2015-torvik

Torvik, V. I., & Agarwal, S. (2016). *Ethnea - An instance-based ethnicity classifier based on geo-coded author names in a large-scale bibliographic database*. Paper presented at the International Symposium on Science of Science, Washington DC, U.S.A. http://hdl.handle.net/2142/88927

Torvik, V. I., & Smalheiser, N. R. (2009). Author name disambiguation in MEDLINE. *ACM Transactions on Knowledge Discovery from Data, 3*(3), 1-29. 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.

Wagner, C. S. (2009). *The new invisible college: Science for development*: Brookings Institution Press.

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:10.1016/j.respol.2005.08.002

Waltman, L. (2012). An empirical analysis of the use of alphabetical authorship in scientific publishing. *Journal of Informetrics, 6*(4), 700-711. doi:10.1016/j.joi.2012.07.008

Wang, D. J., Shi, X., McFarland, D. A., & Leskovec, J. (2012). Measurement error in network data: A re-classification. *Social Networks, 34*(4), 396-409.

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

Watts, D. J. (1999). Networks, dynamics, and the small-world phenomenon. *American Journal of Sociology, 105*(2), 493-527. doi:Doi 10.1086/210318

Yan, E., & Guns, R. (2014). Predicting and recommending collaborations: An author-, institution-, and country-level analysis. *Journal of Informetrics, 8*(2), 295-309. doi:10.1016/j.joi.2014.01.008

Yan, E. J., & Ding, Y. (2009). Applying centrality measures to impact analysis: A coauthorship network analysis. *Journal of the American Society for Information Science and Technology, 60*(10), 2107-2118. doi:10.1002/Asi.21128

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

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

Zhao, D., & Strotmann, A. (2011). Counting first, last, or all authors in citation analysis: A comprehensive comparison in the highly collaborative stem cell research field. *Journal of the American Society for Information Science and Technology, 59*(13), 2070-2086.