Impact of Entity Disambiguation Errors on Social Network Properties

Jana Diesner [1,2], Craig S. Evans [2], Jinseok Kim [1]

Graduate School of Library and Information Science[1], Illinois Informatics Institute[2], University of Illinois Urbana Champaign

Email: jdiesner@illinois.edu, csevans2@illinois.edu, jkim362@illinois.edu

Abstract

Entities in social networks may be subject to consolidation when they are inconsistently indexed, and subject to splitting when multiple entities share the same name. How much do errors or shortfalls in entity disambiguation distort network properties? We show empirically how network analysis results and derived implications can tremendously change depending solely on entity resolution techniques. We present a series of controlled experiments where we vary disambiguation accuracy to study error propagation and the robustness of common network metrics, topologies and key players. Our results suggest that for email data, not conducting deduplication, e.g. when operating on the level of email addressed instead of individuals, can make organizational communication networks appear to be less coherent and integrated as well as bigger than they truly are. For co-publishing networks, improper merging as caused by the commonly used initial based disambiguation techniques can make a scientific sector seem more dense and cohesive than it really is, and individual authors appear to be more productive, collaborative and diversified than they actually are. Disambiguation errors can also lead to the false detection of power law distributions of node degree; suggesting preferential attachment processes that might not apply.

1. Introduction

Social network analysis has become a general utility method for modeling, understanding and explaining patterns and dynamics of social interaction and the mutual influence of (infra-)structure and behavior. One caveat with network analysis is that the validity of results obtained and conclusions drawn heavily depend on the quality and accuracy of the underlying data. When working with electronic records of social interactions, ambiguity of social entities is a challenging key factor in that respect. Errors in ambiguity resolution can be divided into two types:

Firstly, a) the incomplete detection of alternative references to unique individuals, e.g. in the case of spelling variations and synonyms, and b) the incorrect merging of truly distinct entities, e.g. when two or more nodes are represented by the same name string (Hobbs, 1979). The task of correctly identifying and merging all instances of references to the same unique entity is referred to as consolidation. This problem has been studied in natural language processing under the label of co-reference resolution (Bhattacharya & Getoor, 2007; Deemter & Kibble, 2000) and in the context of databases, where it is called record linkage or deduplication (Culotta & McCallum, 2005).

Secondly, failure to correctly split up nodes that truly represent more than one individual (Christen & Goiser, 2007; Sarawagi & Bhamidipaty, 2002). An instance of this effect is when distinct individuals share a common name. For example, according to http://howmanyofme.com/, there are 46,157 John Smith's in the US, and 416 people who have the same name as the authors on this paper. This type of error can also result from erroneous merging, which shows the tight coupling of both types of disambiguation issues. The task of correctly separating nodes that represent multiple distinct entities is referred to as splitting.

While highly accurate algorithmic solutions to consolidation and splitting exist, the impact of deduplication (and if applicable parametric choices) on the robustness of network data, metrics and propagation of errors from data to knowledge are largely unknown. Who cares? We argue that understanding the magnitude and boundaries of variation in network properties that are solely due to disambiguation errors - including not disambiguating data at all - and are thus independent of underlying social processes is essential for the informed planning of data collection and cleaning steps, assessing the legitimacy of results and conclusions, and preventing unjustified further actions, e.g. policy decisions. Based on this lack of knowledge and its significance for network science, we herein address the following question: What impact do errors and improvements in entity deduplication and splitting have on commonly considered node and network properties?

The outcome from this research matters for two more reasons: first, if disambiguation errors had no meaningful impact on network properties, one could spare the costs for this step. Even though this sounds appealing when considering that disambiguation routines may involve some manual inspection, we don't know if this argument is justifiable. Second, disambiguation techniques are less than 100% accurate, and current research in computing focuses on improving these rates. However, the payoff from these incremental improvements is also unknown. In summary, this means that we have a poor understanding of the relationship between minor changes in disambiguation accura-

cy and related changes in networks. The findings presented herein shed some new light on these open questions.

What's next? We review prior work on disambiguation, introduce our datasets (one subject to consolidation, the other to splitting) and experimental design for testing the impact of different types and levels of disambiguation on networks, present our findings, and discuss their meaning and implications.

2. Background

Prior work has shown how different levels of graph incompleteness impact common network metrics, also depending on different topologies (Borgatti, Carley, & Krackhardt, 2006; Frantz, Cataldo, & Carley, 2009). These studies agree in concluding that minor amounts of missing data both on the node and edge level can cause substantial biases in network properties. Understanding these biases is relevant for data collection as this knowledge inform us how much missing data is acceptable. What this line of work leaves unanswered is the question of how ambiguity of data impacts analysis results.

A specific domain where people pay strong attention to entity disambiguation is bibliometrics. There, author-based analyses of publication records have been used to identify influential scholars (Yan & Ding, 2009) and mapping static and evolutionary network properties (Barabási et al., 2002; M. E. J. Newman, 2001). One common challenge for co-authorship and co-citation network studies is the identification of unique authors from digital records, which has been solved in three different ways: Algorithmic solutions to this problem are available (Tang, Zhang, & Yao, 2007) and leveraged by some data providers, e.g. DBLP (Franceschet, 2011). Alternatively, researchers have manually disambiguated author names for studies of a few hundred to thousand publication records (Yan & Ding, 2009). This approach guarantees high-quality disambiguation, but does not scale up. The third disambiguation approach in bibliometrics is the application of heuristic rules. One instance of this approach is to disambiguate people based on the first initial of the given names of authors; a technique also known as first-initial based disambiguation (Liben-Nowell & Kleinberg, 2007; M. E. J. Newman, 2001). For example, if two name instances have the same surname and same first name initial, they are considered to refer to the same person. This technique bears the risk of erroneously merging distinct authors, e.g. in the case of 'Newman, Mark E.' versus 'Newman, Mark W.' A more fine grained version of this approach is to use all available initials of first and middle names, i.e., all-initial based disambiguation (Barabási et al., 2002; M. E. J. Newman, 2001; Radicchi, Fortunato, Markines, & Vespignani, 2009). With this technique, entities are considered identical if they match in the surname and all initials of given name(s). In this case, 'Newman, Mark E.' and 'Newman, Mark W.' would be considered as different entities since they differ in the second initial. This technique can lead to erroneously splitting entities into multiple records when a single person's name appears inconsistently across publications. For example, the one and same 'Newman, Mark E.' may be indexed as 'Newman, M.,' 'Newman, Mark E.,' and 'Newman, M. E. J.'. Despite the potential errors with initial-based disambiguation, this approach has been a prevalent in bibliometrics (Milojević, 2013; Strotmann & Zhao, 2012). Scholars have argued that errors due to this approach are not detrimental to research findings (Barabási et al., 2002; Goyal, van der Leij, & Moraga-Gonzalez, 2006; M. E. J. Newman, 2001), and assume first- and all-initial based disambiguation to provide the lower and upper bound for the true number of authors, respectively.

The issues described for author-based networks generalize within limits to other domains where they are less intensely studied, e.g. people posting and commenting on social media, and participation in collaboration and communication networks. Building upon our prior work (Kim & Diesner, accepted; Kim, Diesner, Kim, Aleyasen, & Kim, 2014), in this paper, we contrast findings for disambiguating co-authorship networks to a specific kind of communication networks, namely email networks.

3. Data and Disambiguation

We use two large-scale, longitudinal datasets that represent a partial view on interactions in real-world social systems. The first one is an email collection that was made available as part of the investigation into Enron. With people having multiple email addresses, this dataset is used to illustrate consolidation impacts. The second dataset is a co-publication dataset of papers available from MEDLINE. Given the standardized name indexing on the latter – and many other - databases, the main issue here is splitting up names that were conflated as multiple people happened to have the same name. The datasets differ in writing style (casual or corporate (Enron) versus scientific (MEDLINE)), intended audience (explicitly specified recipients versus wider scientific audience), mode and frequency of production and delivery (individually generated and distributed in real-time versus co-produced and released over long cycles) and length, but overlap in their coverage of multiple years and size of more than a hundred thousand documents produced by tens of thousands of people (Table 1). We chose these specific datasets because they are subject to consolidation (Enron) versus splitting (MEDLINE). Beyond that, they vary along some dimensions while holding others constant, which helps to scope out the generalizability of our findings.

Table 1: Comparison of Raw Datasets

Characteristic	Enron	MEDLINE
Time Range	10/1999-07/2002	01/2005-12/2009
Number of documents	520,458	101,162
Domain	Email	Co-publishing
Cultural context	Corporate, internal communication	Scientific, external/ public communication

3.1. Email Data Corpus

As part of the investigation into Enron by the Federal Energy Regulatory Commission (FERC), investigators seized computers from the operations in Houston, and the resulting emails were made available to the public. An instance of the raw email dataset was cleaned up and released by William Cohen (2009) and made available for academic research. The raw data is comprised of 150 distinct mailboxes (users), containing approximately 520,000 email messages over a core timespan from October 1999 through to July 2002.

The vast majority of prior network analyses that use the Enron data have created network data by considering email addresses as nodes and email exchange between them as edges. Since people can have more than one email address (we provide facts and statistics on this in the next section), this common procedure introduces duplication errors in the form of redundancies. It also creates ambiguity about the unit of analysis as some nodes represent truly unique individuals, while in other cases, the collective body of information sent or received by a person is spread across multiple nodes.

Each individual email is self-contained in a text file, and largely conforms to RFC822 and RFC2882 text messaging standards. While this sounds like a clean data format, in the early 2000s, there were competing corporate email systems, including Lotus Notes, cc:Mail and Microsoft Outlook. Each implemented the RFC standards in their own way using a combination of the familiar <name>@<domain.com> format and the then common X.400 extended naming standards. During the time period of the emails, from examining the data it is clear that a migration from an X.400 based system (Lotus Notes) to Microsoft Outlook took place.

We parsed each email into three sections: 1) People: a multi-stage effort (outlined below) to match email addresses to actual individuals (deduplication, consolidation). These data are used for building social networks. 2) Body: the text of each message split into a per sentence representation. These data can be used for text mining. 3) Admin: a file containing metadata about the messages (timestamps, file locations, etc.). These sections were loaded into an Oracle database for further analysis.

3.1.1. Ambiguity Resolution: Consolidation

The mapping of email addresses to real people, which was facilitated by using lists of employees in Enron (Diesner, Frantz, & Carley, 2005) is complicated by many users having multiple email addresses, and many people having the same name but different email addresses.

The above mentioned email migration created another challenge where we not only had to deduplicate names and addresses from Outlook, but also deal with issues arising from the conversion from Lotus Notes in which not all email addresses were translated correctly. The impact of these translation inaccuracies are thousands of "orphaned" email addresses with poorly formatted header files.

Consolidating email addresses is a daunting task. While it can be machine assisted with a computer giving a best guess estimate, e.g. based on semantic similarity and associated confidence values, it requires human interaction with the data, including close readings of emails to verify identities. Our disambiguation work has resulted in three instances of the data:

1. **Raw**: raw email addresses, no disambiguation. This includes Enron (@enron.com) and non-Enron (e.g. @anderson.com) addresses. This baseline represents the network data and findings one would obtain if no disambiguation was done.
2. **Disambiguated**: leverages a manually vetted mapping of email addresses to people, including full names, job histories and physical locations (Diesner et al., 2005). Mappings were verified by looking back into the source emails to ensure correctness where needed. This forms the bulk of unique email addresses. Further manual disambiguation of remaining addresses was performed by students as part of a graduate class in data analytics. Non-Enron email addresses are not considered. This represents a "personalized" network where all nodes represent the same unit of analysis, i.e. social entities within an organization.
3. **Scrubbed**: using the disambiguated version plus further manual consolidation and automated scrubbing (explained below). A guided heuristic approach was used where a best guess on the names was created and manually verified to ensure they weren't unbelievable. Over time, we manually verified matches via the original messages. We also removed "garbage" email addresses and mailing lists. This represents an even cleaner dataset and more precisely defined unit of analysis, i.e. actual individuals within an organization.

Stages 2. and 3. comprised a few months of work. The results from this paper give an idea of the return of investment for these efforts. Table 2 shows the number of email addresses per person in the "scrubbed" version. We have identified more than 1 email address for 1,523 people. For those people, the average number of emails per person is 2.4, and the median is 2. Note that without this effort, Kenneth Lay, who served as Chairman, CEO, and Presi-

Table 2: Number of Email Addresses Per Person

No. of email addresses per person	No. of people with that no. of addresses	Person (* indicted)
26	1	Kenneth Lay, Chairman*
11	3	Jeffrey Skilling, CEO* David Delainey, Energy Trader* Vince Kaminski, MD Research
10	3	Susan Scott Steven Kean, EVP, Chief of Staff Mark Haedicke, General Counsel
9	4	Mark Taylor, Asst Gen Counsel Grant Masson, VP Research Patrice Mims Jeff Dasovich, Exec - Gov Affairs
8	5	Too many to name
7	13	
6	17	
5	36	
4	63	
3	160	
2	1,218	
1	21,753	

dent of Enron, for example, would appear as 26 distinct nodes in the network. His email addresses range from well-formed variations of his name such as (kenneth_lay, kenneth.l.lay, kenneth.lay)@enron.com, to contractions of his name {ken_lay, ken.lay)@enron.com, to shortened versions (klay, kllay, layk, ken, lay)@enron.com, to functional, role-based addresses (ken.lay-.chairman.of.the.board, office.chairman)@enron.com. Arguably, he might have used some of his accounts for different types of communication, while others that merely differ in spelling could be considered equivalent in terms of purpose and use. For this paper, we decided not to differentiate between these addresses in order to focus on the effect of proper consolidation.

3.1.2. Network Construction

From our database, we generated three instances of the email network that differ solely in their level of consolidation: raw (worst data quality), disambiguated (better) and scrubbed (best we have). These networks are directed, weighted graphs representing the sender, receiver and number of connections (emails sent) per directed tie.

Table 3 summarizes the impact of the considered consolidation stages on the size of the constructed networks. The deduplication of addresses causes an increase in the number of email addresses per node as we move from a one to one relationship (raw) to many to one in the disambiguated versions. Going from raw to disambiguated, the number of nodes drops sharply (up to 75%, which is also strongly impacted by disregarding non-Enron email addresses). The decreases are more moderate for going from disambiguated to scrubbed (12% to 24% less nodes) as well as on the edge level (36% and 12 %, respectively). In the results section we show how these changes translate into impacts on network properties.

Table 3: Number of Entities and Ties in Email Networks

Number of	Raw	Disambiguated (Diff to Raw)	Scrubbed (Diff to Raw) (Diff to Disamb.)
Senders	19,466	6,205 (-68%)	5,441 (-72%) (-12%)
Receivers	72,713	19,700 (-73%)	15,297 (-79%) (-22%)
Addresses	81,811	20,332 (-75%)	15,526 (-81%) (-24%)
Edges	332,683	212,768 (-36%)	188,045 (-43%) (-12%)

3.2. Co-Publishing Data

MEDLINE is the bibliographic database of the National Library of Medicine. It covers journals in biology and medicine published from 1950 up to now. A baseline version of the MEDLINE data is publicly distributed in XML format every year. The 2012 version contains more than 20 million publication records. Each article is indexed with a unique key identifier (PMID), title, journal name, authors name(s), author affiliation (if available) and medical subject headings (MeSH) assigned by human experts.

To create a dataset of similar size and scope as the Enron corpus, we retrieved a subset of MEDLINE records by querying articles containing the MeSH term 'brain' occurring in publications between 2005 and 2009. This resulted in 109,578 publication records from 3,701 journals. By limiting the search to 'brain' (the most frequent MeSH term from the last 10 years (D. Newman, Karimi, & Cavedon, 2009)), we obtain a topically connected subcommunity – similar to a large international corporation (Enron) with a certain focus. As we are concerned with network properties, single-authored articles were excluded.

3.2.1. Ambiguity Resolution: Splitting

Although each MEDLINE article has a unique identifier, authors are not disambiguated beyond name strings, i.e. a surname is followed by a given name or given name initial(s). Therefore, author names can feature merging or splitting errors. To tackle this problem, we used an Authority database (Torvik & Smalheiser, 2009), where MEDLINE author names were algorithmically disambiguated with high accuracy (up to 98~99%). In that database, each pair of name instances with the same surname as well as first initial of given name in two different articles was compared for similarity using eight features: middle name initial, suffix (e.g., Jr.), journal name, language of article, coauthor name, title word, affiliation word and MeSH term. If 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 (Torvik & Smalheiser, 2009; Torvik, Weeber, Swanson, & Smalheiser, 2005). To obtain disambiguated author names, each article in our dataset was matched with its corresponding article in the Authority database through the unique PMID. We found 101,162 matches from 3,660 journals. The difference to our original sample (109,578 papers, 3,701 journals) is due to the fact that the Authority database uses the 2010 MEDLINE data.

3.2.2. Network Construction

In order the generate the initial based disambiguated networks, each name instance (e.g., Newman, Mark E.) was processed into two versions: (1) surname + first initial of given name (e.g., Newman, M.) and (2) surname + all initials of first and middle names (e.g., Newman, M. E.). Thus, in the resulting networks, an author is represented by a name string in the format of surname plus initial(s). This labeling schema is in sync with the representation of individuals in the disambiguated and scrubbed Enron graphs.

We generated three instances of the co-publishing networks that differ solely in their disambiguation approach: algorithmic (best we have), all initials based (worse yet common) and first initial based (worst we have, also common). A unique author or node is denoted by a unique id based on the Authority database. If two authors appear in the byline of the same paper(s), they get linked. Following prior research (Barabási et al., 2002; Franceschet, 2011; M. E. J. Newman, 2001), only the existence of ties is considered, while their frequency is disregarded. This results in undirected, binary graphs. The algorithmically disambigu-

Table 4: Number of Entities and Ties in Co-Publishing Networks

	Algorithmic	**All-initials (Diff to algorith.)**	**First-initials (Diff to algorith.) (Diff to all-init.)**
Name Instances	557,662	557,662	557,662
Unique Entities	258,971	207,256 (-20.0%)	182,421 (-29.6%) (-12.0%)
Edges	1,335,366	1,317,894 (-1.3%)	1,303,957 (-2.4%) (-1.0%)

ated graph can be considered as a proxy for ground-truth data or the approximately true number of distinct authors.

Table 4 summaries the impact of disambiguation on the size of the constructed co-publishing networks. Merging due to initial based disambiguation (the reverse and actually desired effect is splitting, which can be traced by going from right to left in the table) results in a decrease in the number of nodes by 20% to 30% for all- versus first-initials based disambiguation, respectively. Similar to what was observed for the email graph, the changes are less drastic for going from version two (all-initials) to three (first-initials) and on the edge level.

4. Metrics

Once extracted, the networks were analyzed with R (3.1.1) using the igraph library (0.7.1) and Pajek. The following versions of standard metrics were calculated:

Numbers of Vertices and Edges: cumulative total counts of the number of unique vertices (N) and number of links (binary) connecting them. This indicates communication volume (Enron) and productivity (MEDLINE).

Density: the number of existing edges over the number of possible edges (=N*(N-1)/2).

Clustering Coefficient (CC): the probability of forming an edge between two vertices that have a common neighbor (M. E. J. Newman, 2001), defined as:

$$\mathrm{CC} = 3 \times \frac{\text{number of triangles on the network}}{\text{number of connected triples of vertices}} \quad (1)$$

Shortest Path Length: the lowest possible number of edges connecting a pair of vertices. As nodes in disconnected graphs cannot reach each other, only reachable pairs are considered (Brandes, 2001). Diameter refers to the longest shortest path in a network.

Component: a subset of vertices that all can reach each other. The total number of components and the ratio of the size of the largest component over the total number of vertices are reported.

Centralization (C_A): measures variability or heterogeneity of centrality metrics (A) in networks (Wasserman & Faust, 1994), defined as:

$$\mathrm{C_A} = \frac{\sum_{i=1}^{N}(\max(\mathrm{A}) - \mathrm{A_i})}{\max\sum_{i=1}^{N}(\max(\mathrm{A}) - \mathrm{A_i})} \quad (2)$$

The denominator is the theoretically maximal sum of differences (taken pairwise between vertices) in a centrality A. We calculate centralization scores for six centrality (A) measures (L. C. Freeman, 1977; Linton C. Freeman, 1979):

Degree centrality (A_D): $\mathrm{x_{ij}}$ represents the presence of an edge between vertex i and j without considering directionality. This is calculated for MEDLINE.

$$\mathrm{A_D}(\mathrm{i}) = \sum_{\mathrm{j}=1}^{\mathrm{N}} \mathrm{x_{ij}}(\mathrm{i} \neq \mathrm{j}) \quad (3)$$

In-degree (A_{ID}) and Out-degree (A_{OD}) centrality: consider edge directionality and thus are only calculated for Enron.

$$\mathrm{A_{ID}}(\mathrm{i}) = \sum_{\mathrm{j}=1}^{\mathrm{N}} \mathrm{x_{ji}}(\mathrm{i} \neq \mathrm{j}) \quad (4)$$

$$\mathrm{A_{OD}}(\mathrm{i}) = \sum_{\mathrm{j}=1}^{\mathrm{N}} \mathrm{x_{ij}}(\mathrm{i} \neq \mathrm{j}) \quad (5)$$

Betweenness centrality (A_B): $\mathrm{n_{jk}}$ is the number of shortest paths between vertex i and j, and $\mathrm{n_{jk}}(\mathrm{i})$ is the number of shortest paths between j and k that includes i

$$\mathrm{A_B}(\mathrm{i}) = \sum_{\mathrm{j<k}} \frac{\mathrm{n_{jk}}(\mathrm{i})}{\mathrm{n_{jk}}} \quad (6)$$

Closeness centrality (A_C): $\mathrm{d}(\mathrm{i},\mathrm{j})$ is the shortest path between actor i and other N-1 number of vertices. As this is incalculable for disconnected network, only the largest component of each network is considered.

$$\mathrm{A_c}(\mathrm{i}) = \sum_{\mathrm{j}=1}^{\mathrm{N}} \frac{1}{\mathrm{d}(\mathrm{i},\mathrm{j})} \quad (7)$$

Eigenvector centrality (A_E): λ is a constant and $\mathrm{x_{ij}} = 1$if vertex i is connected to vertex t and $\mathrm{x_{ij}} = 0$ if otherwise (Bonacich, 1972). This is a recursive function of degree.

$$\mathrm{A_E}(\mathrm{i}) = \lambda \sum_{\mathrm{j}=1}^{\mathrm{N}} \mathrm{x_{ij}} \mathrm{A_E}(\mathrm{j}) \quad (8)$$

5. Results

We find that failing to properly merge (Enron) or split (MEDLINE) nodes strongly biases most common network metrics; with distortions on the order of tens to over one hundred percent (all node and network property results are summarized in Table 5). This is true even when changes in the number of links are small (1.3% to 2.4% in MEDLINE). For about half of the network metrics, the changes in values exceed the change in the number of nodes, which is already 20% to 81%. Incremental efforts to mitigate disambiguation issues pay off in terms of moving results closer to true values. The changes in metrics depending on the dataset and disambiguation approach are shown in Table 5.

For Enron, the rawer or less deduplicated the data, the larger and more fragmented the networks appear to be; suggesting less cohesion and more distance among members in an organizational communication network than there truly is. This can lead to false – as in overly negative - conclusions about organizational culture and needs for collaboration tools.

For co-publishing, both commonly applied initial based disambiguation techniques lead to the impression of more dense, connected, cohesive and centralized networks with shorter paths and less components than there really are. Individual authors seem more productive, collaborative and diversely connected. First-initials based disambiguation leads to even stronger biases than the all-initials method. This suggests a more vibrant and integrated science sector and better performing scientists than reality has it.

In summary, each disambiguation step affects network properties to a non-negligible extent. We next explain the causes for these impacts for the example of co-publishing; with the same reasoning - just in the opposite direction - being applicable to the email networks.

The decrease in unique entities from algorithmic (best)

Table 5: Network properties for email and co-publishing networks per disambiguation method (ratio of change to raw/ algorithmic in parentheses)

	Email Networks (directed)			**Co-Publishing Networks (undirected)**		
Network version	Raw Emails	Manual Disambiguation	Scrubbed	Algorithmic Method	All-initials Method	First-initial Method
Data quality	Worst	Better	Best	Best	Worse	Worst
Main effect	Consolidation of nodes increasing (left to right), Elimination of issues (left to right)			Splitting up of nodes increasing (right to left), Introduction of issues (left to right)		
Size	Number of nodes and edges decreasing (left to right)			Number of nodes and edges decreasing (left to right)		
No. of Nodes	81,811	20,332 (-75.15%)	15,526 (-81.02%)	258,971	207,256 (-19.97%)	182,421 (-29.56%)
No. of Edges	332,683	212,768 (-36.04%)	188,045 (-43.48%)	1,335,366	1,317,894 (-1.31%)	1,303,957 (-2.35%)
Density	4.97E-05	5.14E-04 (+9.34%)	7.80E-04 (+14.69%)	3.98E-05	6.14E-05 (+54.27%)	7.84E-05 (+96.98%)
Clustering Coefficient	0.07637	0.09421 (+18.94%)	0.10698 (+28.61%)	0.39	0.20 (-48.72%)	0.19 (-51.28%)
Diameter	18 (Directed) 15 (Undirected)	10 (Directed) 10 (Undirected)	10 (Directed) 7 (Undirected)	22	19 (-13.64%)	18 (-18.18%)
Avg. Shortest Path Length	4.33	3.56 (-17.78%)	3.56 (-17.78%)	6.70	5.21 (-22.24%)	4.78 (-28.66%)
No. of Components	978	10 (-98.98%)	5 (-99.49%)	10,182	5,028 (-50.62%)	3,100 (-69.55%)
Ratio of Largest Component	96.82%	99.91% (+3.09%p)	99.95% (+3.13%p)	80.91%	90.47% (+9.56%*p*)	93.63% (+12.72%*p*)
Degree Centralization	N/A	N/A	N/A	1.83E-03	6.98E-03 (+281.42%)	8.40E-03 (+359.02%)
In Degree Centralization	0.01635	0.03052 (+86.67%)	0.03561 (+117.80%)	N/A	N/A	N/A
Out Degree Centralization	0.01909	0.07858 (+311.63%)	0.07858 (+311.63%)	N/A	N/A	N/A
Eigenvector Centralization	0.99588	0.98552 (-1.04%)	0.98213 (-1.38%)	0.212	0.195 (-8.02%)	0.187 (-11.79%)
Betweenness Centralization	0.01041	0.02014 (+93.47%)	0.02728 (+164.65%)	9.85E-03	2.26E-02 (+129.44%)	2.09E-02 (+112.18%)
Closeness Centralization	N/A	N/A	N/A	0.159	0.228 (+43.40%)	0.238 (+49.69%)
Closeness (In) Centralization	1.14E-05	1.04E-06 (-90.88%)	8.40E-07 (-92.63%)	N/A	N/A	N/A
Closeness (Out) Centralization	1.98E-05	8.19E-04 (+40.36%)	1.98E-03 (+99%)	N/A	N/A	N/A

to all (worse) and first (worst) initial based disambiguation means that many author identities get merged, although splitting also exists (the same happens for Enron as we go from raw (worst) to refined (better and best) data). Edges decrease at a much lower rate because this only happens if not only two authors with identical names are merged, but their coauthors also get consolidated. A rather unlikely scenario, yet it happens.

For density, relatively small changes in the number of edges (numerator) coupled with the strong decrease in nodes (denominator) lead to drastic increases in density due to disambiguation.

The merging of authors also impacts distance-based metrics: as identities get consolidated, merged vertices act as bridges or shortcuts between local networks of merged authors. This results in an overall shrinkage of networks, which reduces nodal distances. This shrinkage explains the decreases in diameter, average shortest path length and number of component. Components that are truly disconnected from the largest one become attached to the dominating component due to the merging effect; leading to an increase in the largest component ratio and overall cohesion.

Centralization measures aim to capture how strongly a network centers around relatively small groups of relatively central nodes; i.e. it expresses the inequality in the distribution of centrality scores across nodes. Incorrectly merged nodes have a higher chance to get unduly more often linked to others, sit more often on shortest paths among others and can access others more quickly than the remaining nodes. Consequently, unjustified merging leads to inflated degree, betweenness and closeness centrality scores for the impacted vertices, which causes increased centralization values on the graph level. Interestingly, eigenvector centralization shows a decrease for both datasets due to disambiguation. This might be explained by the fact that eigenvector scores are high for immediate neighbors of well-connected vertices. The merging of entities produces artifact entities with high centralities; an effect that can then impact their neighbors. Such an overall increase of eigenvector scores of individual vertices seem to contribute to the decrease of the centralization of eigenvector scores across the network.

The only exception to the observed common trends for email and co-publishing graphs is the Clustering Coefficient (CC): as the number of vertices per network decreases due to merging and splitting, network densities increase, but surprisingly, the CC only went up for the email graphs. In co-publishing, erroneous merging distorts the CC; underestimating the tendency of coauthors to collaborate. This is because the co-authors of incorrectly merged authors are unlikely to have co-published with the same people. In contrast, in email data, as vertices are merged, the likelihood of one person having sent/received an email

Table 6: Changing Rank based on Centrality Measures and Corpus Versions

(entries redundant for all three networks per dataset with gray background, entries redundant for two networks in italics)

Degree Centrality						
Rank	**Enron**			**MEDLINE**		
	Raw	**Disambiguated**	**Scrubbed**	**Algorithmic**	**All Initials**	**First Initial**
1	sally.beck@enron.com	Beck, Sally	Beck, Sally	Krause, W	*Wang, Y*	*Wang, J*
2	kenneth.lay@enron.com	*OUTLOOK TEAM*	Lay, Kenneth	Fulop, L	*Wang, J*	*Wang, Y*
3	jeff.dasovich@enron.com	Forster, David	Forster, David	Nawa, H	*Wang, X*	Lee, J
4	david.forster@enron.com	Lay, Kenneth	Jones, Tana	Su, Y	Chen, Y	Kim, J
5	*outlook.team@enron.com*	TECHNOLOGY	Kaminski, Vince	Medarova, Z	Li, X	*Wang, X*
Closeness Centrality						
Rank	**Enron**			**MEDLINE**		
	Raw	**Disambiguated**	**Scrubbed**	**Algorithmic**	**All Initials**	**First Initial**
1	*outlook.team@enron.com*	Lay, Kenneth	Beck, Sally	Trojanowski, JQ	*Wang, J*	*Wang, J*
2	john.lavorato@enron.com	Beck, Sally	Lay, Kenneth	Kretzschmar, HA	*Wang, Y*	*Wang, Y*
3	david.forster@enron.com	*OUTLOOK TEAM*	*Kitchen, Louise*	Toga, AW	*Wang, X*	*Wang, X*
4	sally.beck@enron.com	*Kitchen, Louise*	Kean, Steven	Thompson, PM	Li, X	Lee, J
5	kenneth.lay@enron.com	Lavorato, John	Lavorato, John	Barkhof, F	*Zhang, J*	*Zhang, J*
Betweenness Centrality						
Rank	**Enron**			**MEDLINE**		
	Raw	**Disambiguated**	**Scrubbed**	**Algorithmic**	**All Initials**	**First Initial**
1	jeff.dasovich@enron.com	Beck, Sally	Beck, Sally	Toga, AW	*Wang, J*	*Wang, J*
2	kenneth.lay@enron.com	*Kaminski, Vince*	Lay, Kenneth	Kretzschmar, HA	*Wang, Y*	*Lee, J*
3	sally.beck@enron.com	Lay, Kenneth	*Kaminski, Vince*	Thompson, PM	*Wang, X*	*Wang, Y*
4	gerald.nemec@enron.com	*Skilling, Jeffrey*	Jones, Tana	Trojanowski, JQ	Li, J	*Wang, X*
5	*jeff.skilling@enron.com*	OUTLOOK TEAM	Hayslett, Rod	Barkhof, F	*Lee, J*	Zhang, J
Eigenvector Centrality						
Rank	**Enron**			**MEDLINE**		
	Raw	**Disambiguated**	**Scrubbed**	**Algorithmic**	**All Initials**	**First Initial**
1	louise.kitchen@enron.com	Kitchen, Louise	Kitchen, Louise	Futreal, PA	*Wang, Y*	*Wang, Y*
2	sally.beck@enron.com	Beck, Sally	Beck, Sally	Stratton, MR	*Liu, Y*	*Wang, J*
3	john.lavorato@enron.com	*Haedicke, Mark*	*Haedicke, Mark*	Edkins, S	*Wang, J*	*Liu, Y*
4	david.forster@enron.com	Lavorato, John	Lavorato, John	Omeara, S	*Wang, X*	*Wang, X*
5	tana.jones@enron.com	Forster, David	Forster, David	Stevens, C	Li, X	Zhang, J

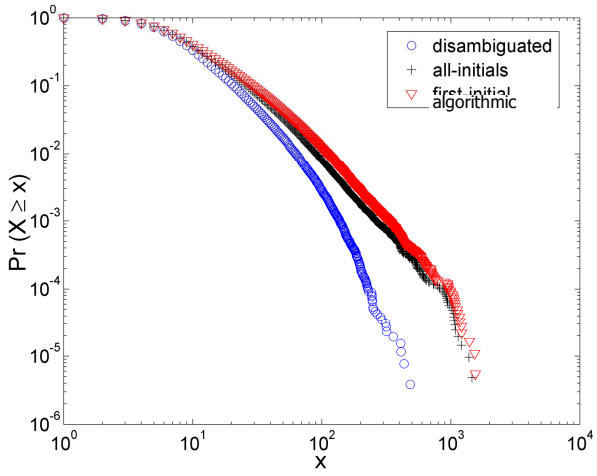

Figure 1: MEDLINE - Cumulative log-log plot of degree distribution per disambiguation method

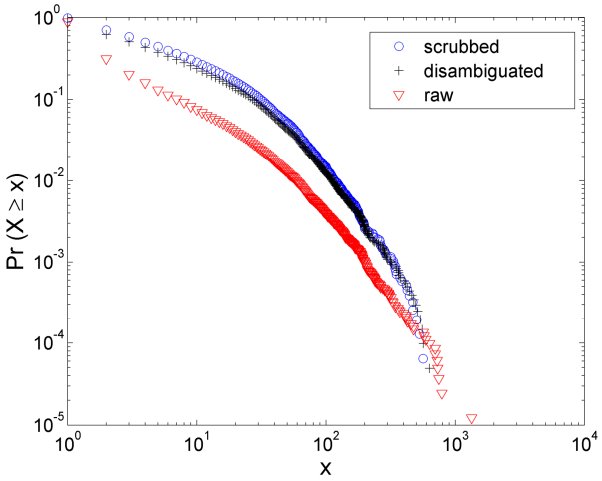

Figure 2: Enron - Cumulative log-log plot of in-degree distribution per disambiguation method

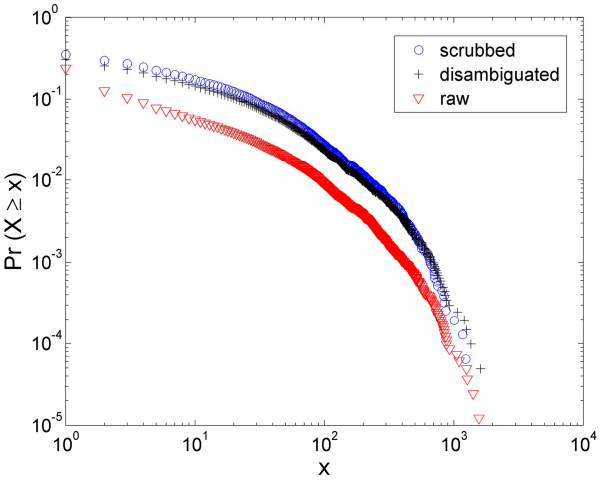

Figure 3: Enron - Cumulative log-log plot of out-degree distribution per disambiguation method

message from the same person as another one increases due to the nature of corporate organizations. Also, this differences in trends for the CC can be explained by the nature of the datasets: in Enron, the data exhibit homophilic tendencies across multiple dimensions (location, business units, organizational role), while ties of this nature are fewer in MEDLINE (different locations and organizations, fewer people sharing these characteristics).

The next logical question here is whether the impact of disambiguation errors on network data is so strong that it can distort our understanding of applicable network topologies. Both communication and coauthorship networks have shown to be following a power-law distribution of node degree; implying that these networks evolve based on preferential attachment of new members to well established or highly popular ones (Barabási et al., 2002; Milojević, 2010; M. E. J. Newman, 2001). Figures 1 to 3 show the degree distributions per disambiguation approach and dataset as cumulative log-log plots. For indegree in the email graph (Figure 2), increasing data accuracy correlates with an increasing tendency to feature a power law distribution, even if such a distribution is valid for a limited number of vertices. We do not observe this effect for outdegree (Figure 3), which seems logical as one person (versus multiple people or a mailing list) can only send so many emails, while communication coming in to a person allows for preferential attachment effects more naturally. In the co-publishing graphs, the algorithmically disambiguated data (Figure 1) show the most curved line, while the lines for all-initial-based disambiguation and even more so first-initial-based disambiguation are straighter. This might lead to falsely assuming that these networks begin to be driven by preferential attachment.

Besides the so far presented quantitative bird-eye view on network properties, we also tested for the impact of entity resolution on the identification of key players based on centrality metrics. The top five authors per dataset and disambiguation techniques in terms of degree, eigenvector, betweenness and closeness centrality are shown in ranked decreasing order in Table 6. We get different pictures depending on the cultural contexts: for Enron, we see a moderate overlap in individuals across versions of the network and also across metrics (mailing lists shown in upper case letters). For example, Sally Beck occurs among the top 5 for all email networks and metrics, and Kenneth Lay in 2/3 of all cases. This suggests that key player analysis is more robust to disambiguation flaws than network metrics. In other words, highly central individuals will still feature prominent in highly incorrect data. In contrast to that, in the co-publishing graphs, there is no overlap between the algorithmically versus initial-based disambiguated graphs, but a strong intersection among the two initial-based methods. This means that the set of top key players in the ground-truth proxy are completely different from those in the initial based disambiguated networks. This effect is mainly due to the strong presence of scholars with common Asian names (mainly Chinese and Korean), such as Wang, Lee and Zhang; leading to strong improper merging effects. This finding implies that accurate disambiguation is essential for data that entail a considerable ratio of Asian names, and that research on name disambiguation needs to take such cultural aspects into consideration.

6. Discussion and Conclusions

We have shown how network analysis results can tremendously differ depending on entity disambiguation techniques and respective errors; possibly leading to erroneous conclusions about network properties, topologies and key players. Similar trends were found for two longitudinal, large-scale datasets from different domains. Our findings separate domain-specific effects from more general impacts of error propagation.

For email data, not conducting deduplication can make organizational communication networks appear to be bigger as well as less coherent and integrated than they truly are. This might lead to false conclusions for managing organizational communication, e.g. stimulating integration through additional meetings or collaboration tools, and overestimating the need for more interaction. For co-publishing networks, improper merging as caused by the commonly applied initial-based disambiguation techniques can make a scientific sector seem more dense and cohesive than it really is, and individual authors appear to be more productive and collaborative than they truly are. This might overestimate the impact of collaboration and funding, and underestimate the need for collaboration.

Key player analysis is more robust to disambiguation errors than the other tested network properties – but only for cultural context where people have somewhat unique combinations of first, middle and last names. In other words, a large ratio of network participants with common Asian names – as typical in a wide range of scientific domains - can also severely distort key player analysis results calculated based on improperly disambiguated data.

In summary, the presented results suggest that entity resolution affects our understanding of macroscopic, graph-level features as well as microscopic investigations of influential network members. We argue that highly accurate disambiguation is a precondition for testing hypotheses, answering graph-theoretical and substantive questions about networks, and advancing network theories.

Who cares about these findings? Even though disambiguation techniques have been developed for specific domains or applications, e.g. (relational) databases of bibliometric records, many of these methods share a large portion of assumptions and data pre-processing approaches. We argue that a better understanding of the impact of entity resolution errors and the robustness of network properties towards these errors contribute to a greater comparability and generalizability of findings. Our findings help to improve the understanding of scalable, robust and reliable methods for constructing social networks from digital records. The gained knowledge can assist researchers and practitioners in drawing valid conclusions from their work and the work of others, and should encourage us to pay closer attention to proper entity resolution. These data provenance issues are particularly relevant when archiving or reusing existing data where it might not be clear what pre-processing techniques have been employed in what way.

One counterargument to our problem definition could be that big data will wash out entity resolution issues, e.g. because consolidation and deduplication errors balance each other out. There are no prior empirical findings for judging the validity of this lame proposition, but the results from this study provide solid arguments against it.

We are currently expanding this work to develop routines for identifying sets of nodes and edges for which consolidation or splitting seems most applicable and impactful for the overall data. This will help users to focus in on a smaller set of nodes with a high return of investment for disambiguation efforts.

Acknowledgment

This work is supported in part by KISTI (Korea Institute of Science and Technology Information, P14033). Some preprocessing for this project was made possible in part by the Institute of Museum and Library Services (RE-05-12-0054-12, Socio-technical data analytics SODA). The disambiguated MEDLINE dataset was provided by Vetle Torvik and Brent Fegley from GSLIS/ UIUC.

References

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

Bhattacharya, I., & Getoor, L. (2007). Collective entity resolution in relational data. *ACM Transactions on Knowledge Discovery from Data, 1*(1), 1-36.

Bonacich, P. (1972). Factoring and weighting approaches to status scores and clique identification. *Journal of Mathematical Sociology, 2*(1), 113-120.

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.

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

Christen, P., & Goiser, K. (2007). Quality and complexity measures for data linkage and deduplication *Quality Measures in Data Mining* (pp. 127-151): Springer.

Cohen, W. W. (2009). Enron Email Dataset. Retrieved 3/17/2014, 2014, from http://www.cs.cmu.edu/~./enron/

Culotta, A., & McCallum, A. (2005). *Joint deduplication of multiple record types in relational data.* Paper presented at the 15th ACM International Conference on Information and Knowledge Management (CIKM).

Deemter, K., & Kibble, R. (2000). On Coreferring: Coreference in MUC and Related Annotation Schemes. *Computational Linguistics, 26*(4), 629-637.

Diesner, J., Frantz, T. L., & Carley, K. M. (2005). Communication Networks from the Enron Email Corpus. "It's Always About the People. Enron is no Different". *Computational & Mathematical Organization Theory, 11*(3), 201-228.

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 & Mathematical Organization Theory, 15*(4), 303-328.

Freeman, L. C. (1977). Set of Measures of Centrality Based on Betweenness. *Sociometry, 40*(1), 35-41. doi: 10.2307/3033543

Freeman, L. C. (1979). Centrality in social networks: Conceptual clarification. *Social Networks, 1*(3), 215-239. doi: 10.1016/0378-8733(78)90021-7

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

Hobbs, J. (1979). Coherence and coreference. *Cognitive science, 3*(1), 67-90.

Kim, J., & Diesner, J. (accepted). The Impact of Data Pre-Processing on Understanding the Evolution of Collaboration Networks. *Journal of Informetrics*.

Kim, J., Diesner, J., Kim, H., Aleyasen, A., & Kim, H.-M. (2014). *Why name ambiguity resolution matters for scholarly big data research.* Paper presented at the IEEE BigData, International Workshop on Challenges & Issues on Scholarly Big Data Discovery and Collaboration Washington DC.

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

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

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

Newman, D., Karimi, S., & Cavedon, L. (2009). Using Topic Models to Interpret MEDLINE's Medical Subject Headings. In A. Nicholson & X. Li (Eds.), *AI 2009: Advances in Artificial Intelligence* (Vol. 5866, pp. 270-279): Springer Berlin Heidelberg.

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

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

Sarawagi, S., & Bhamidipaty, A. (2002). *Interactive deduplication using active learning.* Paper presented at the Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining.

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

Tang, J., Zhang, D., & Yao, L. (2007). Social network extraction of academic researchers. In N. Ramakrishnan, O. R. Zaiane, Y. Shi, C. W. Clifton, & X. D. Wu (Eds.), *Icdm 2007: Proceedings of the Seventh Ieee International Conference on Data Mining* (pp. 292-301).

Torvik, V. I., & Smalheiser, N. R. (2009). Author Name Disambiguation in MEDLINE. *Acm Transactions on Knowledge Discovery from Data, 3*(3). 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: 10.1002/Asi/20105

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

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