Email Alias Detection Using Social Network Analysis

Ralf Hölzer
Information Networking Institute
Carnegie Mellon University
Pittsburgh, PA 15213
rholzer@cmu.edu

Bradley Malin
School of Computer Science
Carnegie Mellon University
Pittsburgh, PA 15213
malin@cs.cmu.edu

Latanya Sweeney
School of Computer Science
Carnegie Mellon University
Pittsburgh, PA 15213
latanya@privacy.cs.cmu.edu

ABSTRACT

This research addresses the problem of correctly relating aliases that belong to the same entity. Previous approaches focused on natural language processing and structured data, whereas in this research we analyze the local association, or "social" network in which aliases reside. The network is constructed from email data mined from the Internet. Links in the network represent web pages on which two email addresses are collocated. The problem is defined as given social network S, constructed from email address collocations, and an email address E, identify any aliases for E that also appear in S. The alias detection methods are evaluated on a data set of over 14,000 University X email addresses for which ground truth relations are known. The results are reported as partial lists of k choices for possible aliases, ranked by predicted relational strength within the network. Given a source email address, a portion of all email addresses, 2%, are correctly linked to another alias that corresponds to the same entity by best rank, which is significantly better than random (0.007%) and a geodesic distance (1%) baseline prediction. Correct linkages increase to 15% and 30% within top-10 (0.07% of all emails) and top-100 rank lists (0.7% of all emails), respectively.

1. INTRODUCTION

Individuals on the Internet use aliases for various communication purposes. Aliases can be tailored to specific scenarios, which allows individuals to assume different aliases depending on the context of interaction. For example, many online users utilize aliases as pseudonyms in order to protect their true identity, such that one alias is used for web forum postings and another for e-mail correspondence. Determining when multiple aliases correspond to the same entity, or *alias detection*, is useful to a variety of both legitimate and illegitimate applications. Regardless of the intent behind alias detection, it is important to understand the extent to which the process can be automated.

When aliases are listed on the same webpage it can indicate there exists some form of relationship between them. In order to leverage this relationship, we analyze several methods for alias detection based on social network analysis [19]. Social network analysis has recently been integrated into the computer science community to model several problems, including record linkage in co-authorship networks [4] and name disambiguation [13]. We assume the network in which aliases, extracted from webpages, are situated reveal certain aspects of the social network to whom the alias corresponds.

Since many people use several email addresses for related purposes, we attempt to determine which email addresses correspond to the same entity by analyzing the relational network of addresses extracted from webpages. Email addresses, a type of alias, can be distilled from a large number of web pages, such as class rosters [18], research papers [5], resumes [12], discussion boards, or USENET message archives [7]. For this paper networks are constructed from email addresses extracted from web pages within a specific university's system. As a result, similarities in the local network surrounding each address can be exploited to determine which aliases correspond to the same entity. Furthermore, email addresses provide another useful property for determining relationships. In contrast to other identifiers, email addresses provide a unique mapping from address to a specific entity. Thus, no disambiguation is necessary when studying email addresses as identifiers for alias detection.

The remainder of this paper is organized as follows. Section 2 reviews earlier approaches to alias detection and determining importance between nodes in social networks. Novel methods based for alias detection are discussed in section 3. In addition, the graph representation of the network and the ranking algorithms are introduced. In section 4, the detection methods are evaluated on a dataset for which a large number of email aliases are known. Results and limitations of the approaches are discussed in section 5.

2. RELATED RESEARCH

Alias detection is related to the problem of alias disambiguation. The latter attempts to determine if the same alias, such as "John Smith", refers to one or multiple entities. There are certain similarities between the disambiguation and detection, and as a result, some of the methods and insights garnered from one can be applied to the other. In this section we review several approaches which have been applied to the disambiguation and detection problems. The approach of choice depends primarily on the type of underlying data to be analyzed.

Natural language processing has been successfully applied to identify whether separate writings have been authored by the same individual. Computational and statistical models were first proposed by Mosteller and Wallace [14] to solve disputes regarding the authorship of free text documents. Their models were extended by Rao et al. [17] who applied techniques from linguistics and stylometry to identify pseudonyms in a textual context on the Internet. These methods were successful in identifying aliases used by the

LinkKDD 2005, August 21, 2005, Chicago, Illinois, USA.

same individual, even if these individuals applied technical measures to disguise their identity. Novak et al. [15] have also developed text analysis algorithms to determine aliases used on the web. These algorithms are based on analyzing the text actively posted on the web under a variety of aliases.

In contrast to free text analysis, several researchers have focused on the analysis of structured bibliographies to characterize authorship. Han et al. [8] used machine learning methods to disambiguate names and pseudonyms in citation data, where an author publishes under similar but not identical names. Similarly, Pasula et al. [16] developed a probabilistic algorithm to solve the problem of ambiguous citations in scientific publications. The latter method is based on formal algorithms for linkage of similar records in databases with defined attributes set forth by Winkler [21]. Yet, these approaches are highly dependent on the structure of information surrounding the entities.

Recently, researchers have turned to social network analysis techniques for alias detection and name disambiguation [4, 3, 9, 10, 11, 13]. Similar to the previous methods, Hill [9] and Hsiung et al. have built classifiers for aliases based on relational networks that were trained in a supervised environment. For instance, Hill constructed classifiers for paper authors that are derived from co-citation data. When provided with a new paper, the author of which was unknown, the citation-based classifiers were used to determine the author. From an unsupervised perspective, Bhattacharya and Getoor [4, 3] extended Winkler's record linkage methods [21] by incorporating co-authorship link structure of the underlying data. These algorithms use an iterative process for deduplication in order to determine if two identifiers refer to the same entity. This approach is similar to alias detection, where two identifiers refer to a single real-world entity. Though this method is tailored to social networks which manifest as clique structures, alternative has been developed for name disambiguation in less centralized social systems [11, 13]. One such approach, proposed by Malin [13], is based on an importance ranking in a relational network surrounding the entity in question. The method looks at collocations and the size of the source from which identifiers are extracted. Unlike the methods proposed in this paper, these network-based approaches fail to explicitly account for the impact of source size and number of collocations independently.

Whereas the previous studies attempted alias detection and disambiguation, Adamic and Adar [1] studied methods to determine relationship importance from mailing lists and other data on the web. Their weighting scheme for predicting similarity in a social network is similar to the weighting algorithms in this paper, but only uses a single, combined measure. White and Smyth [20] have previously developed algorithms to determine importance in social networks in a more general setting. However, these algorithms for determining importance between nodes do not take any heuristics into account.

3. ALIAS DETECTION METHODS

To detect multiple aliases corresponding to the same entity via network analysis, aliases are collected from sources with collocations. In the case of email addresses the sources are web pages listing several email addresses. In this section, we describe network representation and similarity measurements between aliases pairs.

3.1 Data Representation

Let S represent the set of sources from which identifiers are extracted. Let I be the set of unique email addresses, where I_s denotes the subset of addresses listed on a source $s \in S$. The social network of email addresses is modelled as an undirected graph $G = (I, E)$. Each node $i \in I$ is a distinct email address and each edge $e_{ab} \in E$ is a list of sources in which i_a and i_b were collocated. Let $c_{ab} = |e_{ab}|$ denote the number of sources associated with each edge connecting a and b. As a corollary, the network contains an edge between each pair of email addresses that collocated on at least one. Similarly, there exists a clique (i.e. non-null edge) between all addresses on $s \in S$.

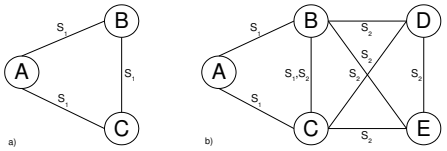

Figure 1: *a)* shows the graph with one source S_1 containing the identifiers $\{A, B, C\}$. *b)* shows the graph after a source S_2 with identifiers $\{B, C, D, E\}$ has been added.

Figure 1a) shows a network containing a single source $S_1 = \{A, B, C\}$, were A, B, and C are identifiers on the source. In Figure 1b), a second source $S_2 = \{B, C, D, E\}$ with the corresponding identifiers has been added. The identifiers B and C are collocated on two different sources, represented by the two sources listed on the edge connecting them.

3.2 Ranking Algorithms

This section describes the ranking methods. A ranking is a top-k list of possible aliases, with the most likely alias candidates at the top of the list. A shortest path algorithm is used to generate a ranking of nodes closest to a given originating node. Nodes closer to the source are favored over nodes at a further geodesic distance.

A useful measure to describe the distance between two nodes in the network is geodesic distance [19]. The geodesic distance between originating node a and destination node b is the length of the shortest path from a to b. The first approach ranks possible aliases for a source alias by geodesic distance in an unweighted network.

The goal of the subsequent ranking methods in this section is to adjust the weights on the edges connecting two nodes according the source data. Specifically, relationship strength is augmented using two heuristics: 1) the number of aliases on a source and 2) the number of collocations of aliases. For these weighting heuristics, the edges have a default weight of 1. Each method reduces the weight to a minimum of $\frac{1}{2}$. All weights are normalized and constrained to the interval $\left[\frac{1}{2}, 1\right]$. For ranking purposes, to preserve a minimum distance on each path, values lower than $\frac{1}{2}$ are not permitted.

3.2.1 Geodesic

Potential aliases are ranked from lowest to highest geodesic distance.

3.2.2 Multiple Collocation

This weighting schema models the assumption that two aliases which collocate on more than one webpage signifies a stronger relationship. For this method, the weights on the graph are calculated as

$$multicol_{ij} = \frac{1 + \prod_{i=1}^{c_{ij}} \frac{1}{2^{i-1}}}{2} = \frac{1}{2} + \frac{1}{2^{c_{ij}}}.$$

The weight is reduced according to an exponential decay function on the number of sources in common. The first source will have a large impact, whereas each additional source will have decreasing impact on the reduction of the weight. This ensures that any additional collocation will be taken into account, but with decreasing impact. As a result, the weight has an upper bound of $c_{ij} = 1$ and has an asymptotic lower bound of $\frac{1}{2}$ with each additional edge, where smaller values represent higher importance.

3.2.3 Source Size

For this weighting schema, we model the belief that the strength between two aliases is inversely correlated with the number of aliases in a source. To evaluate this assumption, edges in the graph are reweighted by setting the weight to

$$sourcesize_{ij} = 1 - \frac{1}{|s_{ij}|}.$$

Since the smallest number of identifiers on a source is 2, the minimum distance for this weight is $\frac{1}{2}$. For large values of $|s_{ij}|$, the weight is asymptotic to 1.

3.2.4 Combined

This approach integrates both of the previous assumptions. Basically, all edges between nodes i and j are weighted using minimum akin to control theory method. The weight is calculated as:

$$combined_{ij} = Max\left(1 - \sum_{i=1}^{c_{ij}} \frac{1}{\alpha \times |s_{ij}|}, \frac{1}{2}\right)$$

Each additional edge reduces the weight by a certain amount dependent on the number of identifiers on the source, such that large source sizes reduce the weight less than smaller source sizes. The resulting weight is on the interval $[\frac{1}{2}, 1]$, where, again, a smaller weight indicates higher importance. Since the weight is reduced for each additional edge and the number of edges is theoretically unlimited, the maximum reduction is upper bounded by $\frac{1}{2}$. For this research, we set $\alpha = 10$, so it takes a maximum of 20 edges with a source size of 2 into account, after which there will be no additional weight reduction.[1]

4. EXPERIMENTAL EVALUATION

This section analyzes the methods described above using email address data derived from Carnegie Mellon University (CMU) web pages. For this analysis, a dataset of CMU-specific email addresses were extracted. This dataset contains 1978 distinct email aliases, with ground truth relations known for all, which makes the dataset amenable to evaluation.

4.1 Data Set Statistics

Due to the way in which email addresses are assigned and can be chosen at CMU, each individual is assigned a unique id in the university-wide andrew.cmu.edu, or *Andrew*, email domain. Many departments have self-maintained email subdomains, which provide additional email addresses for each individual in that department. For example, a graduate student in the Department of Electrical and Computer Engineering (ECE) may use a second email address in the ece.cmu.edu subdomain in addition to Andrew. Since usernames in most of these subdomains correspond to those assigned in the Andrew system, it is possible to generate an accurate list of email aliases from the data set. All addresses that were clearly not a person, such as *root*, *webmaster*, or *cs-students* were removed from consideration. The set of aliases is summarized in table 1.

We found 18%, 45%, and 11% coverage of emails in the database for Andrew, SCS, and ECE email address, respectively. However, the percentages are a rough coverage estimate since some inactive emails no longer in the directory can exist on collected webpages. Since most people in ECE and SCS have an email address in more than one CMU subdomain and therefore have an alias in the dataset, the probability of finding at least one email address for a person in the database is higher than the percentages shown.

Total # of aliases	1978
# of distinct individuals	897
individuals with 2 aliases	767
individuals with 3 aliases	100
individuals with 4 aliases	17
individuals with 5 aliases	6
individuals with 6 aliases	6

Table 1: Aliases in the Carnegie Mellon dataset.

In order to determine if collocated email addresses on webpages provide a foundation for non-random networks, several simple tests were run. To see whether an average individual at Carnegie Mellon can be found on the web (and therefore in the graph), several email directories where compared to the contents in the database. These directories included a full list of all active Andrew email accounts in the university-wide email system, all email accounts in the Department of Computer and Electrical Engineering, and the emails listed in the directory of the School of Computer Science (SCS). Table 2 shows the percentage of emails in these directories that are contained in the graph.

Directory	In DB	# of emails	Percentage
Andrew	38764	6835	18%
SCS	903	2003	45%
ECE	161	1504	11%

Table 2: Percentage of email addresses in database, per directory.

Table 3 shows the path lengths generated by running all-pairs shortest paths on all email addresses in the data set. Intuitively, the average path length between any two email addresses across the entire data set should be higher than the average path length between email addresses in a certain department. The results in Table 3 support this hypothesis. The networks where generated by selecting only those paths that have both a source and the destination address with the corresponding subdomain. The intermediary nodes did not need to be in that specific subdomain.

4.2 Results

The ranking methods described above were then applied to the dataset for evaluation. Several different statistics are presented in this section that support the source size and multiple collocation heuristics.

4.2.1 Geodesic Alias Distances

[1] In the data set analyzed, the maximum number of collocations for two aliases was less than 20 pages. Thus, value of 10 should be adjusted for different data sets where large number of collocations are observed.

Subdomain	Avg.	Max.	Stddev.	emails
All	4.15	12	1.18	14766
cs.cmu.edu	3.76	11	1.20	2897
ece.cmu.edu	3.14	8	1.25	514
cald.cs.cmu.edu	1.70	2	0.63	11
privacy.cs.cmu.edu	2.63	4	0.71	99
speech.cs.cmu.edu	1.82	6	1.60	42

Table 3: Path lengths per subdomain.

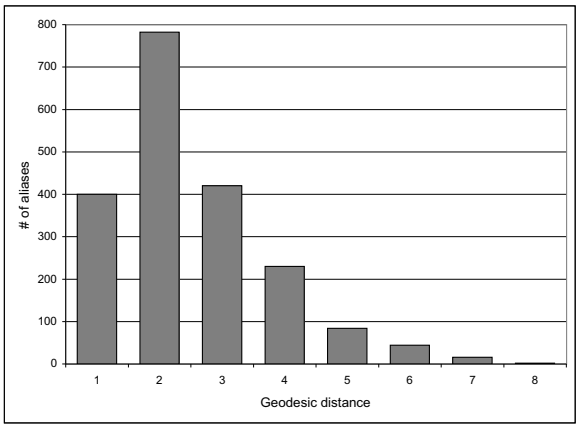

Figure 2: Geodesic distance of all email address pairs.

One hypothesis made earlier was that aliases corresponding to the same entity are close to each other within the network. Figure 2 shows the distances between all pairs of email addresses in the data set. More than 50% of email pairs corresponding to the same entity are within a geodesic distance of two. Moreover, about 400 aliases are within a distance of one. On average, pairs of an entity's email addressed were 2.5 geodesic distance from each other, which is significantly shorter than an average of 4.15 for randomly chosen email pairs. This indicates that pairs of email aliases which correspond to the same entity occur within relatively close proximity of each other.

4.2.2 Method Comparison

Figure 3 summarizes the ranking results. For each known email address in the set of well known aliases, several top-k rankings were generated. These rankings consist of k possible alias candidates, as determined by each method described in Section 3. The top-k lists contain the likely aliases ordered by importance in descending order. The results of each ranking were compared to the set of known aliases. Figure 3 shows the percentage of email address corresponding to the same entity as the source email address found in the top-k results for each alias. The smallest ranking included the top 10 likely alias candidates, whereas the largest ranking included 100 alias candidates. All rankings are out of a total of more than 14,000 email addresses.

It is possible for several email addresses to be ranked identically. Consider, the ranking based on geodesic distance assigns the same rank to all email addresses within the same distance from the source. If several email addresses are tied at the same rank in the results, the median position with the rank is used. For example, if 9 email addresses are tied at rank 1, an alias within these 9 emails would be reported as rank 5.

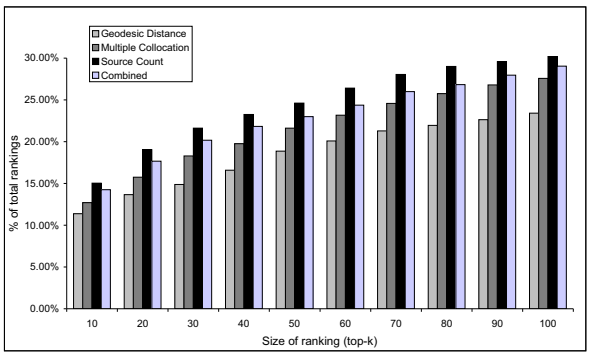

Figure 3: Percent of top-k rankings with at least one email address corresponding to the same entity as the source email address.

Figure 4 shows a comparison of the average precision-recall (PR) curves [2]. The ranking returned by a ranking method can be viewed as the result to a query, where the query is an email address and the result is a ranked list of possible alias candidates. Precision measures the fraction of results in the ranking which are relevant. Recall is the fraction of relevant items in the ranking which have been retrieved.

The PR curves were constructed using the rankings for all emails in the data set that have 6 aliases. The results for all cases have been averaged to produce an average precision and recall curve. Each level of recall represents one of the five aliases for each of these email addresses and therefore measures recall levels from 20% to 100%. Note, the precision for the combined ranking method, which incorporate both source size and multiple collocation heuristics, lies above the baseline (i.e. single heuristic and raw geodesic ranking) approaches for all levels of recall.

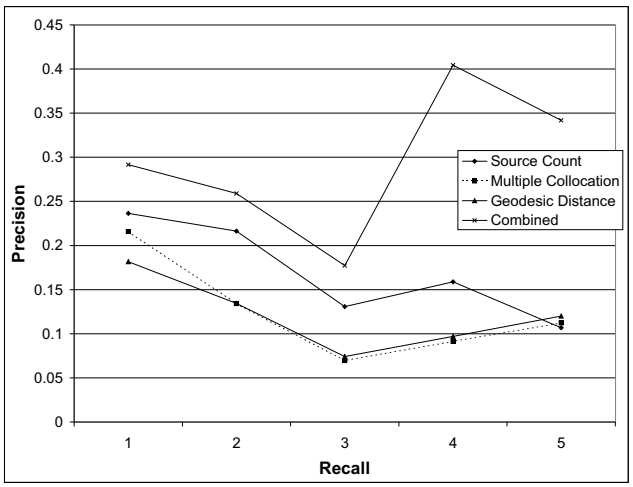

Figure 4: Average precision-recall curves for entities with 6 email addresses.

Figure 5 shows the number of predictions that were correctly identified at rank 1. From this figure it can be seen that all three heuristic methods described above perform better than picking one element from the email addresses at a geodesic distance of one. Specifically, the combined method almost doubles the probability of finding an alias at the first position in the ranking.

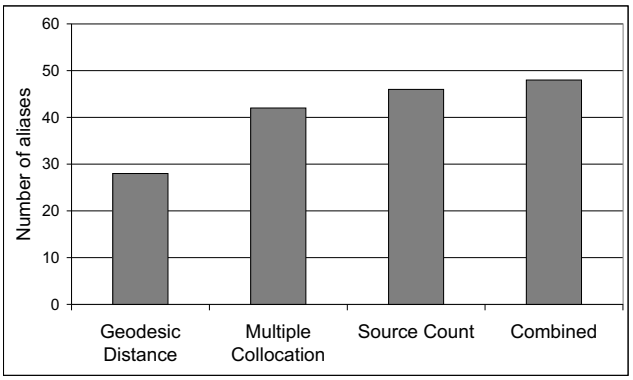

Figure 5: Number of aliases at rank 1.

5. DISCUSSION

This section discusses the results of the experimental evaluation above and addresses certain limitations of the applied methods.

5.1 Findings

The above analyses of email addresses in a social network setting demonstrates that most aliases tend to occur in close proximity of each other. Specifically, more than half of the email aliases are located within a geodesic distance of two from each other and about 20% of aliases are directly connected to each other. This confirms the hypothesis earlier regarding proximity and provides a basis for additional analysis.

The geodesic distance provides a useful method of locating a candidate list of aliases which correspond to the same entity. In more than 10% of the cases, an email alias can be found in the 10 closest addresses in the relational network. The methods developed for determining relationship strength are improvements upon geodesic-based rankings. First, Figure 3 shows that the multiple collocation heuristic is an effective method to increase the probability of finding an additional email aliases corresponding to the same entity. Second, making email address relationships inversely proportional to source size proves even more successful for increasing the probability of finding aliases in the results. Our results demonstrate that taking the number of email addresses on the page into account can increase the probability of finding an alias to more than 15% within the top 10 results, or a 0.0007 fraction of the total email addresses. It is interesting to note that the combined heuristic approach is not as effective as source size only, but does yield results better than geodesic distance.

Figure 5 demonstrates that each method improved the probability of finding an alias. In about 1.1% of the cases, an alias was found at rank 1. The use of the combined approach almost doubled the number of aliases to 2%. Even though the total number of aliases at rank 1 is small, all three methods using a heuristic measure for relationship strength significantly increased this number.

Figure 4 shows an average precision and recall curve for a small subset of aliases. This subset consisted of individuals with six email aliases. For each alias, the precision and recall curve was generated, by determining the rank at which each of the five remaining aliases where found in the total ranking of 14000 email addresses. The combined approach maintains a high level of precision over all levels of recall. The source size method also outperforms the simple geodesic method. Taking multiple collocation into account only showed minimal improvement. These results are mostly consistent with the other results.

5.2 Limitations and Improvements

Empirical results above demonstrate they are feasible in a controlled environment, such as a university, but a more thorough analysis is required. It is unclear how these algorithms will perform in a more general setting, such as the open Internet. One fundamental concern is that it is difficult to obtain a gold standard dataset. Thus unsupervised methods for evaluation must be designed.

Furthermore, there are many extensions to our detection methods which may increase success rate. Here, we briefly address several. First, a portion of the 14,000 email addresses studied correspond to non-human entities. One possible approach to correct this problem is to use a rule-based filter. Simple filter rules for common non-human users, such as “subscribe” or “feedback” may be simple and effective.

latanya.sweeney@cmu.edu
latanya@andrew.cmu.edu
latanya@cs.cmu.edu
latanya@lab.privacy.cs.cmu.edu
latanya@privacy.cs.cmu.edu

Table 4: Examples of email aliases with common id strings.

Second, usernames studied in the dataset are shared across the different domains, making it possible to determine each alias. Many individuals have multiple email addresses that share a common user id part. Table 5.2 depicts various email addresses for Latanya Sweeney in the Carnegie Mellon dataset. Note, though the subdomain changes, the string “latanya” is common to all email addresses. We do not expect that full names will remain constant across email addresses for the same entity, but we do expect there to be logical similarities. Along these lines, Bhattacharya and Getoor [4, 3] demonstrated that string comparator metrics [6], derived from the record linkage community, are feasible for relating name variants in social networks. As a result, we suspect that a comparison of the user id part of the email addresses in the ranked results would make it possible to determine a larger number of correct aliases.

6. CONCLUSION

This research demonstrated that email aliases corresponding to the same entity occur in close geodesic proximity within social networks inferred from online sources. While Geodesic distance provides a large candidate set of email addresses for a source email address, we show ranking methods can discover more precise sets by accounting for source size. Our results suggest that small numbers of email addresses collocated on the same web page are the most likely to have the strongest relationships. The alias detection methods correctly detect a significant number (i.e. better than random) of email addresses using only social relations. Though our methods are limited in precision at best rank predictions, we believe that improvements can be achieved through the incorporation of string comparator similarity metrics and rule-based filters.

7. ACKNOWLEDGEMENTS

The authors wish to thank Benoit Morel, as well as the members of the Data Privacy Laboratory for useful discussions and comments on this research.

8. REFERENCES

[1] L. Adamic and E. Adar. Friends and neighbors on the web. *Social Networks*, 25(3):211–230, 2003.

[2] R.A. Baeza-Yates and B.A. Ribeiro-Neto. *Modern Information Retrieval*. ACM Press / Addison-Wesley, 1999.

[3] I. Bhattacharya and L. Getoor. Deduplication and group detection using links. In *Proceedings of the ACM Workshop on Link Analysis and Group Detection (LinkKDD-2004)*, 2004.

[4] I. Bhattacharya and L. Getoor. Iterative record linkage for cleaning and integration. In *Proceedings of the ACM SIGMOD Workshop on Research Issues in Data Mining and Knowledge Discovery*, 2004.

[5] F. Ciravegna, A. Dingli, D. Guthrie, and Y. Wilks. Integrating information to bootstrap information extraction from web sites. In *Proceedings of the IJCAI Workshop on Information Integration on the Web*, Acapulco, Mexico, 2003.

[6] W. Cohen, P. Ravikumar, and S. Fienberg. A comparison of string distance metrics for name-matching tasks. In *Proceedings of the International Joint Conference on Artificial Intelligence*, Acapulco, Mexico, 2003.

[7] L. Cranor and B.A. Lamacchia. Spam! *Communications of the ACM*, 41(8):74–83, 1998.

[8] H. Han, L. Giles, H. Zha, C. Li, and K. Tsioutsiouliklis. Two supervised learning approaches for name disambiguation in author citations. *Proc. ACM/IEEE Joint Conf on Digital Libraries*, 2004.

[9] S. Hill. Social network relational vectors for anonymous identity matching. In *Proceedings of the IJCAI Workshop on Learning Statistical Models from Relational Data*, Acapulco, Mexico, 2003.

[10] P. Hsiung, A. Moore, D. Neill, and J. Schneider. Alias detection in link data sets. In *Proceedings of the International Conference on Intelligence Analysis*, McLean, VA, 2005.

[11] D. Kalashnikov, S. Mehotra, and Z. Chen. Exploiting relationships for domain-independent data cleaning. In *Proceedings of the 2005 SIAM International Conference on Data Mining*, pages 262–273, Newport Beach, CA, 2005.

[12] N. Kushmerick, E. Johnston, and S. McGuinness. Information extraction by text classification. In *Proceedings of the IJCAI Workshop on Adaptive Text Extraction and Mining*, Seattle,WA, 2001.

[13] B. Malin. Unsupervised name disambiguation via social network similarity. In *Proc. SIAM Wksp on Link Analysis, Counterterrorism, and Security*, pages 93–102, Newport Beach, CA, 2005.

[14] F. Mosteller and D.L. Wallace. *Inference and Disputed Authorship: The Federalist*. Addison-Wesley, Reading, MA, 1964.

[15] J. Novak, P. Raghavan, and A. Tomkins. Anti-aliasing on the web. *Proceedings of the ACM World Wide Web Conference*, 2004.

[16] H. Pasula, B. Marthi, B. Milch, S. Russell, and I. Shpitser. Identity uncertainty and citation matching. *Proceedings of Neural Information Processing Systems*, 2002.

[17] J.R. Rao and P. Rohatgi. Can pseudonymity really guarantee privacy? *Proceedings of the USENIX Security Symposium*, pages 85–96, 2000.

[18] L. Sweeney. Finding lists of people on the web. *ACM Computers and Society*, 34(1), 2004.

[19] S. Wasserman and K. Faust. *Social Network Analysis*. Cambridge University Press, New York, NY, 1994.

[20] S. White and P. Smyth. Algorithms for estimating relative importance in networks. *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2003.

[21] W.E. Winkler. Matching and record linkage. In B.G. Cox, editor, *Business Survey Methods*. Wiley, New York, NY, 1995.