Alias Detection in Malicious Environments

Patrick Pantel

Information Sciences Institute
University of Southern California
4676 Admiralty Way
Marina del Rey, CA 90292
pantel@isi.edu

Abstract

Alias detection is a challenging task critical in several areas such as the intelligence community, social network analysis, databases, biology, and marketing. Problem domains can be as simple as datasets containing accidentally replicated data, or as complex as populations containing criminals or terrorists wielding multiple identities. Teasing out aliases or near aliases in the later case is a serious and challenging problem. We propose an unsupervised information theoretic approach for automatically detecting aliases in malicious environments by observing the behaviors of the entities. Our model discovers the most informative observations (e.g. emails, phone calls, relational data) between entities and then compares them to identify entities exhibiting similar behaviors. We test our model by applying it to the task of discovering aliases in a standard synthetic world of interrelated individuals. Given our system's top-20 guesses, we extracted with 80% accuracy the true aliases of a given entity.

Introduction

Entity consolidation is the problem of uncovering duplicate or near-duplicate entities in a dataset. Problem domains can be as simple as datasets containing accidentally replicated data, or as complex as populations containing criminals or terrorists wielding multiple identities. Teasing out duplicate or near duplicate entities in the later case is a serious and challenging problem.

Alias problems are commonly encountered in the intelligence community when performing background checks or, in general, when tracking individuals from a broad population. Often, simple orthographic cues indicate an alias, as in *Osama bin Laden* and *Usama bin Laden*. Other times, semantic variations can be detected, as in for example *Richard Fendlebaum* and *Dick Fendlebaum*.

Malicious individuals, however, can easily fool such verifications by assigning completely different labels to their identities. But, their *behaviors* are likely to be similar since these are much harder to fake or separate across identities. Behaviors can be observed from various sources of information such as communications (emails, phone calls, chats), transaction material (financial transactions, travel logs, shipments), social links, and other relation data.

For large populations, the total number of such observations can become enormous, with only a small portion of *important* observations overlapping for aliases. Consider the following scenario of a population of Southern California residents and two particular residents *John Doe* and *Alex Forrest*. If you were told that last year both *John* and *Alex* called the Hollywood area about 21 times a month, then would this increase your confidence that *John* and *Alex* are the same person? Yes, certainly. Now, suppose we also told you that *John* and *Alex* both called Kabul about 21 times a month. Intuitively, this observation yields much more evidence that *John* and *Alex* are similar or aliases since not many Southern California residents call Kabul with such frequency. Measuring the relative importance of such observations and leveraging them to detect aliases is the topic of this paper. We will outline an information theoretic framework that models the importance of observations by capturing the intuition of the above example. We will evaluate our system on the task of discovering aliases in a standard artificial dataset created by *Information Extraction & Transport, Inc.* within the EAGLE Project.

The remainder of this paper is organized as follows. In the next section, we review previous approaches to entity consolidation. We then present five important problems that can be cast as an alias detection problem. Subsequently, we outline our information theoretic model and present experimental results. Finally, we conclude with a discussion and future work.

Related Work

Most previous solutions to entity consolidation search for morphologic, phonetic, or semantic variations of the labels associated with the entities. One of the earliest approaches, patented in 1918 by Margaret O'Dell and Robert Russell, was a rule-based system that matched labels which are roughly phonetically alike. This algorithm, later refined as the Soundex matching algorithm [9], removes vowels and represents labels with six phonetic classifications of human

speech sounds (bilabial, labiodental, dental, alveolar, velar, and glottal).

Recently, researchers have looked at combining orthographic (and phonetic) features with semantic features. In addition to string edit distance features, [2] and [7] began looking at the behavioral observations that we introduced in the previous section. They asserted connections between entities for each interrelation present in a link dataset, ignoring the actual relation types. Adding these semantic cues outperformed previous methods like Soundex. Unlike these approaches, our technique makes use of the link labels (e.g. relation types such as *email*, *financial transaction*, *travel to*, etc.) Also, our method automatically determines the importance of each link and leverages this measurement to dramatically reduce the search space.

Davis et al. [5] have proposed a supervised learning algorithm for discovering aliases in multi-relational domains. Their method uses two stages. High recall is obtained by first learning a set of rules, using Inductive Logic Programming (ILP), and then these rules are used as the features of a Bayesian Network classifier. In many domains, however, training data is unavailable. Our method is completely unsupervised and requires no positive or negative samples of aliases. Also, ILP does not scale well to large datasets, whereas our approach does.

A related problem in the natural language processing community is automatic spelling correction. The most widely used systems are based on Shannon's noisy channel model [3][8]. The systems assume that the word that was meant to be written was altered by some corruption model (the noisy channel). A decoder is then trained on tagged examples to reconstruct the original (intended) word given the surface error.

Another related but different problem is when multiple entities are referred to by the same label [11]. For example, the name Michael Jackson refers not only to the singer, but also to the bank president, the talk show host, and the author of several books about beer. This problem is important in natural language applications, such as question answering, which must answer questions such as "Where is the Taj Mahal?" and must select between candidate answers such as Agra, India or Atlantic City, NJ. Most approaches to name resolution have used clustering techniques over coreference chains [1], multiple syntactic and semantic features [10], and over referents by first applying a maximum entropy model that estimates the probability that two labels refer to the same entity [6].

Five Important Aliasing Problems

Solving the aliasing problem is important for many different purposes in various areas such as the intelligence community, social network analysis, databases, biology, and marketing. In some cases, it can be used to flag malicious intents while in others it can be used to clean data and to link knowledge. In this section, we describe five problems that exhibit an aliasing problem at the core: identity thefts, identifying and monitoring terrorist cells, data integration, social network analysis, and author identification. For each of these problems, we describe the entities of the population as well as some of the behavioral observations that may be available for modeling in our system.

Identity Theft

Identity theft is the fastest growing financial crime in the U.S. According to a Federal Trade Commission survey, it is estimated that 27 million Americans have been victims of identity theft in the past five years. In one of its incarnations, thieves acquire social security numbers and other personal information in order to fraudulently acquire credit cards, bank accounts and cell phone accounts. It can be years before unsuspecting victims discover that their credit is ruined, they owe large sums of money to creditors, or worse that they are prosecuted for financial frauds. Supposing authorities are tracking a known identity thief, we expect that our system will be capable of discovering the identities that were stolen by modeling their usage behaviors. The population consists of identities (e.g. all social security numbers) and examples of observed behaviors may include communications, financial transactions, travel information, etc.

Another type of identity theft is when a criminal fraudulently acquires identities, but then sells them instead of using them himself or herself. Here, we expect that our method would not work since the identities are used in different ways by different people.

Terrorist Cells

In 2004, the FBI estimated that al-Qaida sleeper cells were believed to be operating in 40 states, awaiting orders and funding for new attacks on American soil. It is also approximated that between 2,000 and 5,000 terrorist operatives currently operate in the U.S. In response, several researchers developed techniques for threat group detection.

Generally, it is believed that the social links as well as the similar religious and cultural background of terrorists yield cells of very similar people. Treating each terrorist cell as an identity, one can perform group detection to identify its members using aliasing models. Here, the population consists of, say, people living in America and known terrorist cells. The observed behaviors may include communications, financial transactions, travel information, etc. We expect that certain observations will be rare and some will be erroneous, but our model is quite tolerant to sparsity and noise.

Data Integration

In many large organizations and especially government, data is split in many different ways and is collected at different times by different people. The resulting massive data heterogeneity means that staff cannot effectively

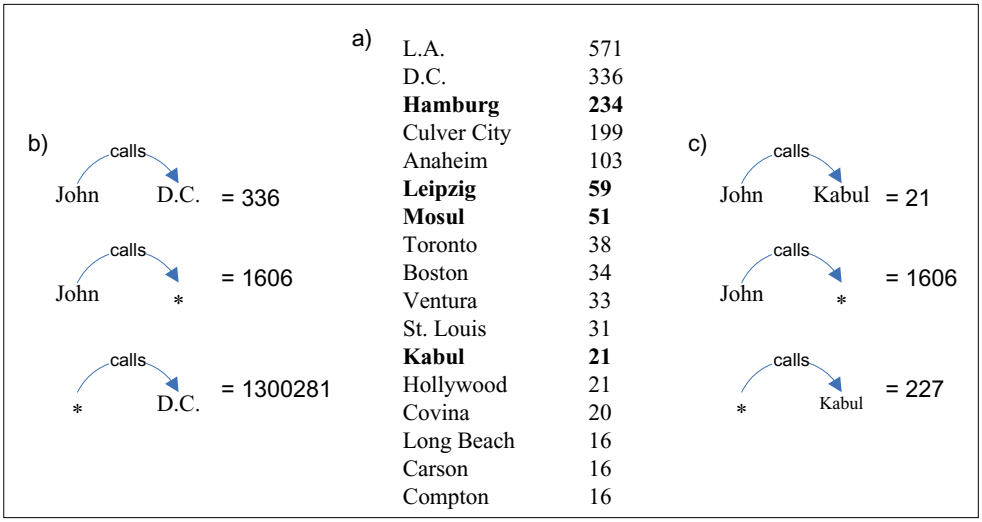

Figure 1. Identifying important observations in our fictitious scenario of phone calls placed by Southern California residents. a) Frequency of phone calls placed monthly by *John Doe*. b) Frequency of calls placed by *John* and others (*) to D.C. and other cities (*). c) Frequency of calls placed by *John* and others (*) to Kabul and other cities (*).

locate, share, or compare data across sources, let alone achieve computational data interoperability.

What is needed is a method to integrate the data in two comparable but heterogeneous data sources. We can partially address this problem by finding matching records across sources. The population consists of all possible records in two data sources, and the observable behaviors are the data fields that are contained in each record. For example, given records of contact names along with contact information such as phone number, address, zip code, etc., we can use this contact information as the observations for our model.

Social Network Analysis

User modeling and recommendation systems are easily hampered by duplicate users in social networks. Duplicate identities, whether accidentally created or not, are prominent. The population here consists of all identities in the network and the observed behaviors include personal attributes such as contact information, colleges attended, board memberships, etc., as well as communication links.

Author Identification

Author identification is the task of attributing authorship to anonymous text [16]. The problem exists in many forms like identifying plagiarism, accrediting famous historical writings to their true authors, and even ownership or copyright legal battles. In the context of alias detection, techniques for author identification can be used to model a source's communication style by looking at the very words and expressions she uses. In an electronic world where one's population consists solely of citizens of the Internet, or Netizens, then one can simply observe the language use and style to find aliases.

Information-Theoretic Model

Most entity consolidation problems consist of large numbers of entities and observations. However, as illustrated in the introduction, certain observations are much more important than others when matching entities (e.g. calling Kabul vs. Hollywood). In this section, we outline an information theoretic model, inspired from the one developed in [12] for aligning databases, that measures this importance.

Before we formally describe the model, we appeal to the reader's intuition. Recall our phone call scenario from the introduction where we were asked if Southern California residents *John Doe* and *Alex Forrest* are the same person given their monthly phone call records. Figure 1 a) lists *John*'s most frequent phone calls along with the call frequencies. It is not surprising that a Californian would call L.A., Culver City, Anaheim, and even D.C. If *Alex* had similar calling patterns to these four cities, it would certainly increase our confidence that him and *John* are the same person, but obviously our confidence would increase much more if *Alex* called the more *surprising* cities Kabul and Mosul.

Looking only at the frequencies of the calls in Figure 1 a), however, one would put more importance on matching calls to L.A. than to Kabul. The goal of our framework is to have a better measurement than frequency for the importance of each call and to re-rank them in order of information content. Figure 1 b) illustrates the frequencies of *John* calling *D.C.*, *John* calling any city, and anyone calling *D.C*, whereas c) illustrates the same for *Kabul*. Notice that although D.C. is much more frequent than Kabul, many more people in the population call D.C. than

Kabul. Our model leverages this observation by adding importance for a city to which *John* calls frequently and by deducting importance if many people in the general population call the same city. After applying our model, the cities in Figure 1 a) are re-ranked as follows:

Kabul	**7.88**	Toronto	4.36
Mosul	**7.05**	Boston	4.31
Leipzig	**5.78**	Covina	2.91
Hamburg	**5.58**	Compton	2.86
Culver City	5.48	St. Louis	2.40
D.C.	5.33	Long Beach	2.03
L.A.	4.77	Carson	1.62
Anaheim	4.46	Hollywood	1.43
Ventura	4.38		

The four cities that were bold in Figure 1 b) are now at the top of this list, and consequently more importance is now put on matched calls to Kabul than matched calls to Hollywood. We now formally introduce our model.

Pointwise Mutual Information

Pointwise mutual information is commonly used to measure the association strength between two events [4]. It essentially measures the amount of information one event gives about another. For example, knowing that a Southern Californian calls L.A. is not informative since most residents call L.A. Conversely, if he or she calls Kabul, then this is an informative observation. The pointwise mutual information between two events x and y is given by:

$$mi(x,y) = \log \frac{P(x,y)}{P(x)P(y)}$$

Pointwise mutual information is high when x and y occur together more often than by chance. It compares two models (using KL-divergence) for predicting the co-occurrence of x and y: one is the MLE (maximum-likelihood estimation) of the joint probability of x and y and the other is some baseline model. In the above formula, the baseline model assumes that x and y are independent. Note that in information theory, mutual information refers to the mutual information between two random variables rather than between two events as used in this paper. The mutual information between two random variables X and Y is given by:

$$MI(X,Y) = \sum_{x \in X} \sum_{y \in Y} P(x,y) \log \frac{P(x,y)}{P(x)P(y)}$$

The mutual information between two random variables is the weighted average (expectation) of the pointwise mutual information between all possible combinations of events of the two variables.

For each entity in our population e, we first construct a frequency count vector $C(e) = (c_{e1}, c_{e2}, \ldots, c_{em})$, where m is the total number of features (observations) and c_{ef} is the frequency count of feature f occurring for entity e. Here, c_{ef} is the number of times we observed feature f for entity e. For example, in Figure 1 b), one feature for e = *John Doe* is f = *Kabul* with frequency 21.

We then construct a mutual information vector $MI(e) = (mi_{e1}, mi_{e2}, \ldots, mi_{em})$ for each entity e, where mi_{ef} is the pointwise mutual information between e and feature f, which is defined as:

$$mi_{ef} = \log \frac{\frac{c_{ef}}{N}}{\frac{\sum_{i=1}^{n} c_{if}}{N} \times \frac{\sum_{j=1}^{m} c_{ej}}{N}}$$

where n is the number of entities and $N = \sum_{i=1}^{n} \sum_{j=1}^{m} c_{ij}$ is the total frequency count of all features of all entities.

In our example from Figure 1 b), assuming that $N = 1.32 \times 10^{12}$, the mutual information for e = *John Doe* and feature f = *D.C.* is:

$$mi_{ef} = \log \frac{\frac{336}{1.32 \times 10^{12}}}{\frac{1{,}300{,}281}{1.32 \times 10^{12}} \times \frac{1606}{1.32 \times 10^{12}}} = 5.33$$

and for f = *Kabul*:

$$mi_{ef} = \log \frac{\frac{21}{1.32 \times 10^{12}}}{\frac{227}{1.32 \times 10^{12}} \times \frac{1606}{1.32 \times 10^{12}}} = 7.88$$

A well-known problem is that mutual information is biased towards infrequent entities/features. We therefore multiply mi_{ef} with the following discounting factor [13]:

$$\frac{c_{ef}}{c_{ef}+1} \times \frac{\min\left(\sum_{i=1}^{n} c_{ei}, \sum_{j=1}^{m} c_{jf}\right)}{\min\left(\sum_{i=1}^{n} c_{ei}, \sum_{j=1}^{m} c_{jf}\right)+1}$$

Similarity Model

Now that we have a method of ranking observations according to their relative importance, we still need a comparison metric for determining the likelihood that two entities are aliases. The requirement is that the metric handles large feature dimensions and that it not be too sensitive to 0-valued features. That is, the absence of a matching observation is not as strong an indicator of dissimilarity as the presence of one is an indicator of similarity. Some measures, like the Euclidean distance, do not make this distinction. Many models could apply here; we chose the cosine coefficient model [14]. The similarity between each pair of entities e_i and e_j, using the cosine coefficient of their mutual information vectors, is given by:

$$sim(e_i, e_j) = \frac{\sum_f mi_{e_i f} \times mi_{e_j f}}{\sqrt{\sum_f mi_{e_i f}^{\ 2} \times \sum_f mi_{e_j f}^{\ 2}}}$$

This measures the cosine of the angle between two mutual information vectors. A similarity of 0 indicates orthogonal vectors whereas a similarity of 1 indicates identical vectors. For two very similar elements, their vectors will be very close and the cosine of their angle will approach 1.

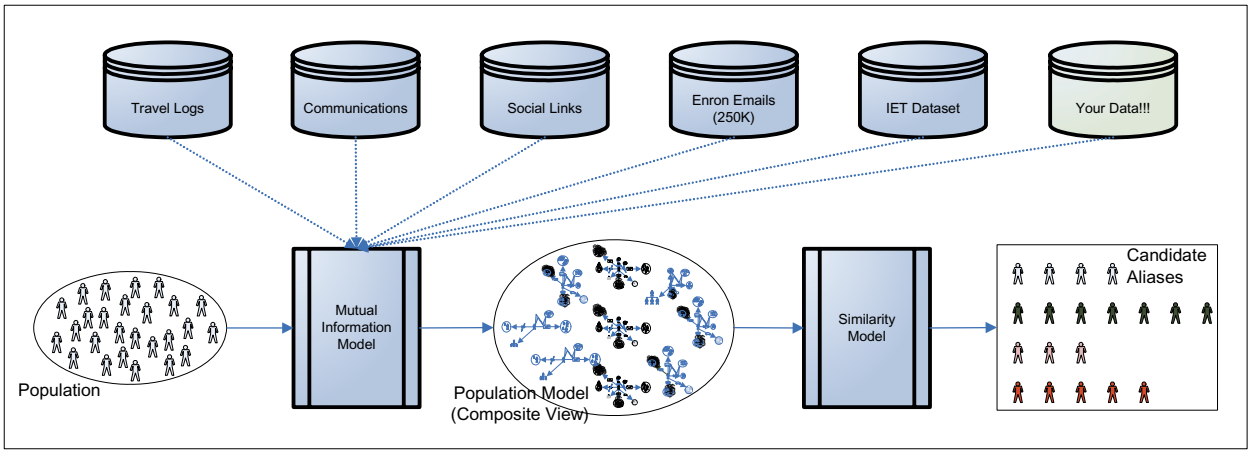

Figure 2. System architecture. First, the mutual information model is applied to the observations of a population. Then, the similarity model yields candidate aliases for each entity in the population.

Detecting Aliases

We now have all the pieces of the puzzle to detect aliases from a given population and a set of observations. Figure 2 illustrates our system architecture for alias detection. The various observations (e.g. travel logs, communications, social links, etc.) are first processed through our mutual information model to generate a ranked composite view of the important observations. Then, our similarity model is used to detect and rank candidate aliases for each entity in our population.

Experimental Results

We evaluate our model on one of the five problems described earlier in this paper: discovering malicious aliases in a social network. Below, we describe our experimental setup and present our results.

Experimental Setup

Following Davis et al. [5], we evaluated our system on standard synthetic datasets modeling threat groups in a social network of a general population, generated by *Information Extraction & Transport, Inc* (IET) within the EAGLE Project [15]. Using an artificial world simulator producing interrelated entities (through communication links and group membership links), IET generated two datasets containing aliases[1]. Each dataset was generated by setting modeling parameters, the most important of which are described below:

- **Population size**: Determines the number of entities in the social network.
- **Connectivity (Sparsity)**: Determines the mean number of interconnections between individuals.
- **Noise**: Determines the signal to noise ratio (omitted connections); the lowest ratio is bounded by: *Easy* (0.08), *Fair* (0.008), and *Hard* (0.0008).

Table 1 summarizes the two IET datasets used in our evaluation. IET provided us with answer keys containing the aliases for each of the two datasets.

Precision and Recall

We applied our model from Section "Information-Theoretic Model", using the communication links as the observations (features) of the entities, and output candidate aliases for each dataset. The precision of the system is the percentage of correct detections:

$$Precision = \frac{C}{T_A}$$

where C is the number of correctly detected aliases and T_A is the total number of proposed aliases by our system. The recall is the percentage of aliases in the answer key, T_M, which were retrieved by the system:

$$Recall = \frac{C}{T_M}$$

Precision and recall measure the tradeoff between identifying aliases correctly and retrieving all of them. Figures 3 and 4 shows the Precision vs. Recall curves of our system on each dataset. The 95% confidence intervals were all in the range of 0.5% – 3%.

Our system shows vastly different performance results on the datasets. Upon inspection of the output, the connectivity and the ratio of total aliases to population size appeared to be the primary cause. Our framework is designed to handle very large feature spaces, which is common in many real world scenarios, by teasing out the most informative features. However, in small feature

[1] IET actually generated five alias datasets for various purposes, but we only focus on the two uncorrupted datasets.

Dataset ID	Population Size	Total Aliases	Connectivity	Noise
18	2504	821	43.6	Fair
20	12,391	681	19.2	Fair

Table 1. Summary of IET datasets for alias detection.

spaces, mutual information cannot as easily uncover the informative features. To support this, we inspected the mutual information vectors for incorrectly identified aliases in dataset 20 and noticed that the mutual information scores are ranked almost identically to the frequency scores.

Increased connectivity is only slightly responsible for the apparent superior performance on dataset 18. It is the ratio of total aliases to population size that had the most drastic impact on the system performance. This is mostly due to the nature of the IET datasets, which were built to model threat and non-threat group behaviors and consequently contained many similar entities (from group memberships). Even though there are few aliases, our system discovers the closeness between entities of same groups. Therefore, the smaller the ratio of aliases to population size is, the lower the system precision will be. In the next section, we show using a different measure that the system performs equally well on both datasets when finding the alias of a suspected entity (i.e. given a particular suspect entity, our system does well at identifying its alias or aliases).

Davis et al. [5] also presented their evaluation on three IET datasets, but they did not specify which IET datasets they used. A direct comparison is therefore impossible. However, like us, they showed vastly different precision and recall curves for the different datasets. Comparing their best precision/recall graph with our precision/recall graph for IET-18 shows that we achieve very similar performance, even though unlike their approach, ours is completely unsupervised and requires no training data.

Top-K Accuracy

In the previous section, we showed that our system sometimes confuses aliases and similar group members (because the IET datasets initially served for the task of group detection). Below, we evaluate our system on the task of identifying the correct alias for a particular suspected entity.

We randomly sampled[2] entities known to have aliases from the IET answer keys, and we measured the percentage of its aliases that were discovered by our system. Figure 5 illustrates the results for varying numbers of Top-*K* guesses (i.e., the system is allowed to make up to

[2] The random samples where taken in order to obtain 95% confidence bounds for our reported accuracies.

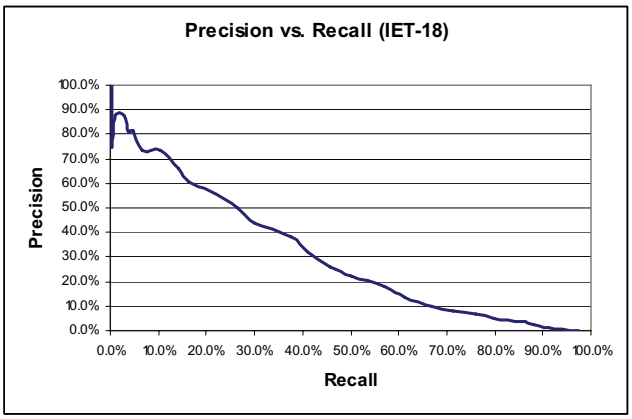

Figure 3. Precision vs. Recall curve for IET-18.

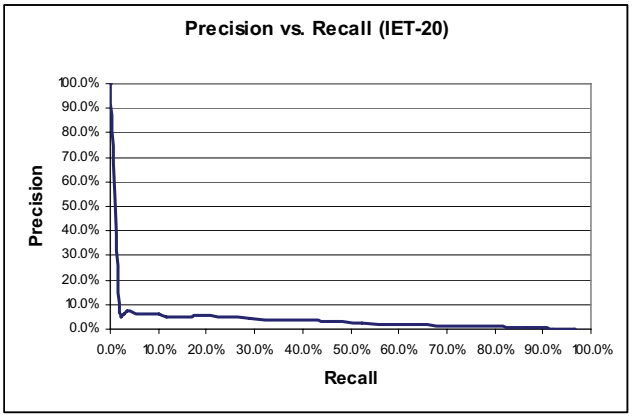

Figure 4. Precision vs. Recall curve for IET-20.

K guesses for each entity, and we evaluate the accuracy of the correct alias being in this list of guesses). Given five guesses, the system finds the correct alias nearly 60% of the time, whereas given 20 guesses the system finds the correct alias 80% of the time. An analyst could potentially use this system for greatly reducing the number of candidate aliases that must be considered in an investigation.

Conclusion and Future Work

Entity consolidation is a common problem encountered in various areas from the intelligence community to social network analysis, databases, biology, and marketing. Instead of detecting aliases by looking for morphological, phonetic, or semantic cues in entity labels, we focused our attention on behavioral cues exhibited by the entities (e.g. communications, financial transactions, social links, etc.), which are much harder to fake in malicious environments. For large populations, the total number of such observations can become enormous, with only a small portion of the *important* observations overlapping for aliases. In this paper, we proposed an information theoretic model for measuring this importance and leveraging it to detect aliases.

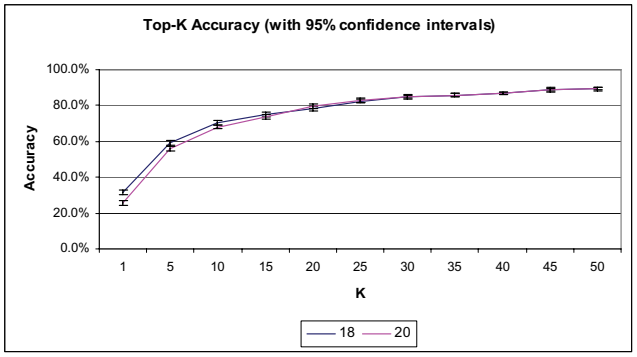

Figure 5. System accuracy on retrieving correct alias in Top-K returned candidates.

We applied our model to the task of discovering aliases from a standard synthetic model of a large, sparse, and noisy social network. Below is a summary of the evaluation results:

- on our best dataset, with 50% accuracy, our system was able to extract 27% of the aliases;
- for a given entity and the top-5 candidate aliases, our system correctly identified the true alias with 60% accuracy;
- for a given entity and the top-20 candidate aliases, our system correctly identified the true alias with 80% accuracy.

At a minimum, our model can dramatically reduce the time a human needs to find the aliases of an entity in a social network (looking at only 20 possible aliases for each entity would uncover the aliases with 80% accuracy). However, the power of the model is critically dependent on gathering the right observations that aliases might share, which in itself is a very interesting avenue of future work. Given the right observations, our model has the potential to solve several serious and urgent problems such as terrorist detection, identity thefts, and data integration.

References

[1] Bagga, A. and Baldwin, B. 1998. Entity-Based Cross-Document Coreferencing Using the Vector Space Model. In *Proceedings of COLING/ACL-98*. Montreal, Canada.

[2] Baroni, M.; Matiasek, J.; and Trost, H. 2002. Unsupervised discovery of morphologically related words based on orthographic and semantic similarity. In *Proceedings of the Workshop on Morphological and Phonological Learning of ACL/SIGPHON-2002*. pp. 48-57. Philadelphia, PA.

[3] Brill, E. and Moore, R. C. 2000. An Improved Error Model for Noisy Channel Spelling Correction. In *Proceedings of ACL-00*. pp. 286-293. Hong Kong.

[4] Church, K. and Hanks, P. 1989. Word association norms, mutual information, and lexicography. In *Proceedings of ACL-89*. pp. 76–83. Vancouver, Canada.

[5] Davis, J.; Dutra, I.; Page, D.; and Costa, V. S. 2005. Establishing Identity Equivalence in Multi-Relational Domains. In P*roceedings of the International Conference on Intelligence Analysis*.

[6] Fleischman, M. B. and Hovy, E. H. 2004. Multi-Document Person Name Resolution. In *Proceedings of the ACL Workshop on Reference Resolution*. Barcelona, Spain.

[7] Hsiung, P. 2004. *Alias Detection in Link Data Sets*. Technical report CMU-RI-TR-04-22, Carnegie Mellon University.

[8] Kernighan, M.D.; Church, K.; and Gale, W. 1990. A pelling correction program based on a noisy channel model. In *Proceedings of COLING-90*. pp. 205-211. Helsinki, Finland.

[9] Knuth, D. 1973. The Art of Computer Programming – Volume 3: Sorting and Searching. Addison-Wesley Publishing Company.

[10] Mann, G. and Yarowsky, D. 2003. Unsupervised Personal Name Disambiguation. In *Proceedings of CoNLL-2003*. Edmonton, Canada.

[11] Martinich, A.P. 2000. *The Philosophy of Language*. Oxford University Press. Oxford, UK.

[12] Pantel, P.; Philpot, A.; and Hovy, E.H. 2005. An Information Theoretic Model for Database Alignment. In *Proceedings of the Conference on Scientific and Statistical Database Management (SSDBM-05)*. pp. 14-23. Santa Barbara, CA.

[13] Pantel, P. and Lin, D. 2002. Discovering word senses from text. In *Proceedings of SIGKDD-02*. pp. 613–619. Edmonton, Canada.

[14] Salton, G. and McGill, M.J. 1983. Introduction to Modern Information Retrieval. McGraw Hill.

[15] Schrag, R. 2004. *EAGLE Y2.5 Performance Evaluation Laboratory Documentation Version 1.5*. Internal report, Information Extraction and Transport, Inc.

[16] Stamatatos, E.; Fakotakis, N.; and Kokkinakis, G. 2001. Computer-based Authorship Attribution without Lexical Measures. *Computers and the Humanities*, Volume 35, Issue 2, May 2001. pp. 193-214.