

# Social Network Relational Vectors for Anonymous Identity Matching

**Shawndra Hill**  
NYU Stern School of Business  
44 W 4<sup>th</sup> Street  
New York, NY 10012  
shill@stern.nyu.edu

## Abstract

Anonymous fraudulent behavior can generate substantial financial burden and inconvenience. Moreover, the recent threat of terrorist infiltration to both business and government has yielded heightened interest in anonymous identity matching (AIM). Most applications of AIM require sophisticated methods robust to issues such as deliberate variation in identity attributes, missing data, and multi-source data corpora. We consider relational social network behavior, eliminating the reliance on personal identifiable data for identity matching. In particular, we consider problems that can be characterized by personal communication networks. We evaluate a proposed implementation of a social network vector-space relational model for AIM on CiteSeer, a research publication citation database.

## 1 Introduction

AIM has garnered attention by government agencies in the wake of perceived increased domestic asymmetric threat. However, civilians are concerned about the potential exploitation of data collected by government agencies, web enabled click stream technologies, credit card companies, and health care providers. The ongoing debate surrounding the tension between security and privacy has motivated data mining research in data privacy.

At the forefront of data mining privacy research are methods that solely rely on perturbed datasets while maintaining predictive performance of various modeling techniques (Agrawal and Srikant 2000; Clifton 2000; Agrawal and Aggarwal 2001) (Atallah, Bertino et al. 1999). More relevant to the AIM discussion are database inference techniques which utilize multi-source data, (Moskowitz 1999; L. Sweeney 2002) to identify individuals who otherwise could not be categorically linked using isolated data sources.

In general, privacy research considers three distinct categories: 1) basic storage and retrieval, i.e., who can access sensitive data; 2) pattern discovery, i.e., the misuse of sensitive data for pattern discovery; and 3) combination of group

patterns, i.e., who can make inferences about individual identity from aggregated data sources (Piatetsky-Shapiro 1995). Despite efforts to encrypt sensitive information, this research indicates that relationship networks may be a subtle indicator of identity.

In this paper, we consider a straightforward method, a social network (Wasserman and Faust 1994; Scott 2000) vector-space model, for AIM in networks of interpersonal relationships. Social network analysis is an appropriate basis for relational learning because: (1) it quantifies relationships; (2) it is well defined; (3) it can be used as a complement to other methods; and (4) it can be used for visualization to enable further understanding of underlying phenomena.

An *actor* is the social entity of interest in a social network. Actors are discrete individuals, or collective social units. In our context, actors may be individuals, companies, industries or nations; we first consider one-mode networks where the actors are considered the same type. A *relational tie* establishes a linkage between a pair of actors. Examples of relational ties include online communication, business transactions, belonging to the same professional club, or a physical/virtual connection. Each actor pair relationship is given a *weight* to indicate the strength. Each actor may have multiple relationships to multiple actors. A vector of weights then represents each actor.

This paper considers research in progress on AIM. We demonstrate the usefulness of the social network vector-space relational model on the application of author identification. The paper is organized as follows. First, we present the social network vector-space model in section 2. In section 3 we apply the vector-space model to the task of author identification and present preliminary results on the CiteSeer database. Finally, we conclude by offering a discussion of results and future research directions in section 4 and 5 respectively.

## 2 Method

For AIM, we would like to classify new relational examples given a set of labeled relational training examples. We consider social network graphs of relationships by

reducing the social network relational graphs to feature vectors of entities. Each new entity in turn represents a candidate example for identification. Weighted term vectors represent all individual entities.

**Definition:** An entity  $e_i$  can be described by an entity vector,

$$e_i = (w_{i1}, w_{i2}, \dots, w_{in}) \quad [1]$$

where  $w_{ik}$  is the weight assigned to the entity  $e_k$  in entity relationship  $e_i$ .

The feature vectors of entities are weighted to give emphasis to stronger entity pair relationships. The weight is determined by the aggregation of all relationships between two entities.

Furthermore, one can specify to what distance in the graph, related entities are considered. At distance one, entities simply represent the weighted adjacency matrix of the relationship graph. At greater distances, however, the entity is composed of embedded entities. To consider entities that embed distant entities, each entity is recursively joined with each of its related entities.

**Definition:** Under addition, entity  $E$  is defined by the weighted union of all nodes (entities), and edges (weights) in  $e_1$  and  $e_2$  where  $\alpha$  and  $\beta$  are scalars.

$$E = \alpha e_1 \oplus \beta e_2 \quad [2]$$

The scalars are utilized to indicate relative significance to the resultant entity. For example, one may want to decay the impact of joined edges in the relationship vector as the distance from the node in the graph increases.

The weight of an edge in  $E$  is therefore defined by [3].

$$w(E) = \alpha w(e_1) \oplus \beta w(e_2) \quad [3]$$

**Definition:** An entity that takes into consideration relational links of distance greater than one may therefore be defined as the recursive sum of each entity with its feature vector entities  $e_k$ .

$$E = \bigoplus_{i=1}^k e_i \quad [4]$$

During the AIM process, new entities are compared to labeled entity vectors. Candidate match sets of entity vectors closest to the query considered similar are ranked and returned. For this exposition, we measure similarity by the cosine distance between the corresponding vector pairs [5]. However, any vector based similarity measure may be considered. The distance measures may be used in

standard hierarchical clustering techniques such as dendrograms (Duda, Hart et al. 2001).

$$dist(q_j, e_k) = \frac{\bar{q}_j \cdot \bar{e}_k}{|\bar{q}_j| |\bar{e}_k|} = \frac{\sum_{i=1}^n w_{i,j} w_{i,k}}{\sqrt{\sum_{i=1}^n w_{i,j}^2} \sqrt{\sum_{i=1}^n w_{i,k}^2}} \quad [5]$$

### 3 Experiment

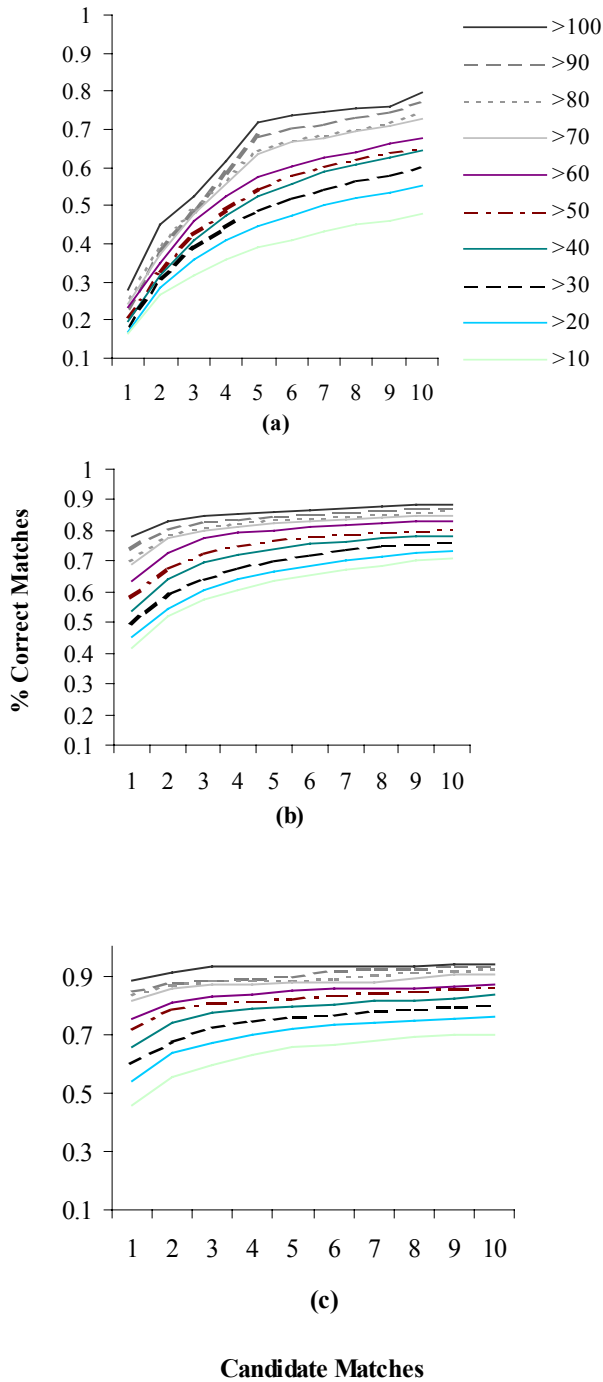
Many refereed journals maintain that anonymity in publication submission is an ethical prerequisite of paramount importance. Nonetheless, we find that reference lists alone identify authors remarkably well. This experiment considers the question of whether the author of a new paper can be identified utilizing solely the citation graph of the paper. We apply the social network vector-space relational model to the CiteSeer database (Lawrence 1999), a scientific literature digital library. We identify authors of papers published in the year 2000 by considering only their citation graph.

Prior Pubs	2000 Pubs	<2000 Pubs	2000 Authors
10	13,174	93,831	8,615
20	9,405	68,597	3,334
30	6,797	50,294	1,659
40	4,678	35,223	855
50	3,462	26,010	510
60	2,636	19,158	315
70	1,932	13,827	191
80	1,540	10,461	128
90	1,201	8,118	91
100	852	5,777	59

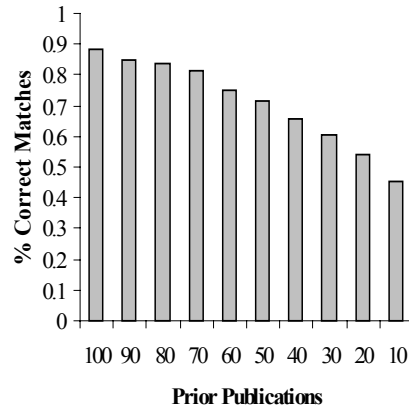
**Table 1:** CiteSeer Data: *Distribution of papers authored and authors with at least n prior publications.*

First, background knowledge is constructed using prior publication knowledge. For each document published prior to 2000 an edge is created linking each author to each cited author. A weighted vector of cited authors defines an author. Next, weighted adjacency vectors are created for each document in 2000. An edge is created between the document and each cited author. A weighted vector of cited authors defines each document. The weights are defined as the total sum of out going links for each author-author, document-author pair.

For this experiment, we are interested in exact author matches as opposed to finding similar authors. Therefore, we limit the dataset to include only papers authored by writers with publication history. The number of prior publications in the background knowledge database determines publication history. We present results for different levels of publication history [Table1] to



**Figure 1:** Author Match Success: *Observed proportions of author matches versus the set size of ranked candidate matches. Each line represents results for documents whose authors had greater than  $n$  publications prior to the year 2000, (a) documents compared to entire CiteSeer database and documents compared to data set segmented by publication history (b) without and (c) with a filter for strength  $>1$*



**Figure 2:** Author Match Success. *Observed proportions of author matches for top ranked candidate versus prior publication record of at least one author.*

understand further attributes that may influence identity matches in large relational networks.

We present results on three experiments motivated by subscription fraud in relational networks (Cortes, Pregibon et al. 2002). We use author identification in the CiteSeer database as a proxy problem to subscription fraud detection.

First, subscription fraud cases generate financial burden to organizations when left undetected. Therefore, the most prudent of methods generate risk scores for all subscribers. Each potential subscriber is compared to the knowledge base of all known subscription fraud offenders before services are rendered. As such, we consider matching documents in 2000 to the entire historical database.

We find that authors with more than 10 prior publications can be identified with 17% accuracy (recall that this is from a total of 8615 authors); 28% of the time the true author is in the top-10 candidate matches. Authors with more than 90 prior publications can be identified with 58% accuracy; 80% of the time the true author is among the top-10 candidates [Figure 1a].

Second, subscription fraud often considers “guilt by association”. In this case, new subscribers are compared only to a subset of fraudulent entities in the knowledge base population that are related in some way. We model this problem by considering sub samples of the knowledge base corresponding to publication history.

We find that authors with more than 10 prior publications can be identified with 78% accuracy; 41% of the time the true author is in the top-10 candidate matches. Authors with more than 90 prior publications can be identified with 87% accuracy; 71% of the time the true author is among the top-10 candidates [Figure 1b].

Finally, a naïve method to remove uninformative edges is to limit the citation graph by considering only relationships with relatively high strength greater than  $n$  [Figure 1c]. The % Correct Matches significantly increased by refining

our search. For authors with more than 10 prior publications, we compared 13,174 documents to 93,831 documents with 8,615 authors [Table 1] and yielded 45.6 % matches to the top ranked candidate [Figure 2]. In practice, filters are used to reduce the time and space complexity of identity retrieval techniques.

In summary, we first show that our simple method works for author matches under different conditions for both the knowledge base and test set. In an attempt to further refine our search and reduce noise in the knowledge base, we followed with an experiment utilizing smaller samples of the knowledge base segmented by publication history. This task refinement resulted in a significant increase in author identification from 28% to 41% in task accuracy with more than 10 prior publications. Finally, we attempt to reduce noise in the test set by filtering less informative nodes which in turn yield accuracy of 45 % on authors with greater than 10 publications in the past. It is important to note that results are shown for different candidate set sizes because in practice, human operators often have the ability to work multiple cases.

## 4 Discussion

In this paper, we introduce the concept of social network vector space model for anonymous identity matching. We concentrate on the method and show preliminary results on a real world citation database.

Our results indicate that considering the network structure of author's reference list does remarkably well at identifying authors, and combining the social network vector-space model with (for example) linguistic analysis may perform even better (Khmelev 2000).

If we can further understand relationships between research community/author network identification and fraud detection, we may inform subscription fraud identification techniques with our method where test labels are abundantly available.

## 5 Future Research

There are many interesting challenges, to behavioral AIM. First, personal communication networks are dynamic and require data structures (Cortes, Pregibon et al. 2002) that capture network evolution through time. Furthermore, the strength of a relationship may not always be determined by absolute frequency. A less "frequent" relationship may be a stronger indication of identity. In general, communication networks are large but sparse. Techniques are needed to preserve graph structure while reducing dimensionality. AIM techniques must consider that communication networks are inherently noisy, fraudulent individuals for example may either attempt to hide their identity or steal that of someone else. Finally, evaluation methods are needed to assess unlabelled anonymous entities matches

The research synopsis considers research in progress. In the future, we will consider multi-attribute entity relationships in our model. We will consider complement-

ing the AIM ranking with available identifiable actor information. In addition, we plan to add linguistic analysis attributes to our relationship vector in the future for author identification. We want to further develop a data structure that will incorporate the dynamic nature of communication links.

We will compare and contrast AIM results of other vector space models such as naive Bayes, information retrieval TF-IDF, and support vector machines. Furthermore, we will demonstrate the efficacy of the proposed method to other communication network domains such as web logs, email logs, long distance calling records and prepaid calling card records. Finally, we plan to investigate appropriate evaluation methodologies where test labels are non-existent.

## Acknowledgements

Thanks to Foster Provost for valuable comments on this draft and continued research support. Daryl Pregibon and Corinna Cortes for research funding in multi-relational data mining. Data used in this paper were drawn from the CiteSeer database (<http://citeseer.nj.nec.com/>). This work is sponsored in part by the Defense Advanced Research Projects Agency (DARPA) and Air Force Research Laboratory, Air Force Materiel Command, USAF, under agreement number F30602-01-2-585.

## References

- Agrawal, D. and C. Aggarwal (2001). On the design and quantification of privacy preserving data mining algorithms. In Proceedings of the Twentieth ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems, Santa Barbara, California, USA, ACM.
- Agrawal, R. and R. Srikant (2000). Privacy-preserving data mining. In Proceedings of the 2000 ACM SIGMOD International Conference on Management of Data.
- Atallah, M. J., E. Bertino, et al. (1999). Disclosure limitation of sensitive rules. Proceedings of IEEE Knowledge and Data Engineering Workshop, Chicago, IL.
- Clifton, C. (2000). "Using Sample Size to Limit Exposure to Data Mining." Journal of Computer Security 8(4).
- Cortes, C., D. Pregibon, et al. (2002). Communities of Interest for Dynamic Graphs. In the Proceedings of Knowledge Discovery and Data Mining Conference, Edmonton, Canada.
- Duda, R. O., P. E. Hart, et al. (2001). Pattern classification. New York, Wiley.

Khmelev, D. V. (2000). "Disputed Authorship Resolution through Using Relative Empirical Entropy for Markov Chains of Letters in Human Language Texts." Journal of Quantitative Linguistics 7(3): 201-207.

L. Sweeney, L. (2002). "K-anonymity: A model for protecting privacy. International." Journal on Uncertainty, Fuzziness, and Knowledge-based Systems 10(7): 557-570.

Lawrence, S., Giles, L., Bollacker, K (1999). "Digital Libraries and Autonomous Citation Indexing NEC Research Institute." IEEE Computer 32(6): 67-71.

Moskowitz, L. C. a. I. S. (1999). Parsimonious Downgrading and Decision Trees Applied to the Inference Problem. The Workshop of New Security Paradigms.

Piatetsky-Shapiro, G. (1995). "Knowledge Discovery in Personal Data versus Privacy---A Mini-Symposium."

Scott, J. (2000). Social network analysis : a handbook. London ; Thousand Oaks, Calif., SAGE Publications.

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