The Myth of the Double-Blind Review? Author Identification Using Only Citations

Shawndra Hill
New York University
44 W 4th Street, 8th Floor
New York, NY 10012
shill@stern.nyu.edu

Foster Provost
New York University
44 W 4th Street, 8th Floor
New York, NY 10012
fprovost@stern.nyu.edu

ABSTRACT

Prior studies have questioned the degree of anonymity of the double-blind review process for scholarly research articles. For example, one study based on a survey of reviewers concluded that authors often could be identified by reviewers using a combination of the author's reference list and the referee's personal background knowledge. For the KDD Cup 2003 competition's "Open Task," we examined how well various automatic matching techniques could identify authors within the competition's very large archive of research papers. This paper describes the issues surrounding author identification, how these issues motivated our study, and the results we obtained. The best method, based on discriminative self-citations, identified authors correctly 40-45% of the time. One main motivation for double-blind review is to eliminate bias in favor of well-known authors. However, identification accuracy for authors with substantial publication history is even better (60% accuracy for the top-10% most prolific authors, 85% for authors with 100 or more prior papers).

Keywords

KDD Cup competition, author identification, social network analysis, relational learning, vector-space model, discriminative self-citations

1. INTRODUCTION

The peer review process, applied to the publication of academic journal articles as well as conference submissions, has been held as the premier control mechanism for the quality of scholarly publications [1]. In most cases, the journal editorial board and a select group of anonymous qualified experts jointly determine the fate of a proposed research article on the basis of scholarly contribution. Although most authors agree the peer review process is indispensable [2], the process is not without flaws [3].

The main problems have been elaborated in the medical field, where the dissemination of low-quality information has the potential to influence loss of life. Among the many concerns [4] is bias [5]. Authors may be discriminated against based on their affiliation and demographic characteristics or preferred because of their reputation or influence in the scholarly community. One potential way to address bias is the adoption of a double-blind review process.

In a single-blind review process the author does not know the identity of the reviewers. In a double-blind review process the identities of both the authors and the reviewers are concealed. A journal may adopt a double-blind process to improve fairness or to improve the perception of fairness (or both). (There is evidence that referees are more critical when they are unaware of the authors' identities [6].)

The American Journal of Public Health, a journal that practices double-blind reviews, surveyed 312 reviewers in 1989 to identify author and institution of reviewed manuscripts. The results indicate that authors could be identified by reviewers using the combination of the paper's reference list and the referee's personal background knowledge [7].

As part of the KDD Cup 2003 competition, we analyzed a very large archive of physics papers. Our goal was to assess how well authors of papers could be identified using only the citations made in the papers. We examined several methods for automatic identification, falling into two general classes: (1) a (dynamic) vector-space model that represents both papers and author histories, and (2) tallying (discriminative) self-citations. The self-citation based methods generally worked better. However, the vector-space models are able to match (with much lower accuracy) even when self-citations are removed.

With the best method, based on discriminative self-citations, authors can be identified 45% of the time. Additionally, the top-10% most prolific authors can be identified 60% of the time. Extremely prolific authors can be identified much more often; for example, authors with 100 or more prior publications can be identified 85% of the time.

Author identification by citation matching is directly related to social network analysis [8], graph theory [9], and bibliometrics [10]. These research areas all study similar methods for graph matching. In social network analysis, the task of identity matching is cast as a structural equivalence problem [11], in standard graph theory, as subgraph isomorphism, and in bibliometrics, as bibliographic coupling. Historically, bibliographic coupling [12] has been used to establish the subject similarity of documents for information retrieval. Text-based methods also have been used for author attribution [13, 14] and for gender classification [15].

We include a *dynamic* vector-space model because publication records have a temporal dimension. For example, authors move to new research areas and change citation behavior as they progress through a research program. Hence, an author's citation history is at best only an approximation of future citation behavior. Inexact graph matching is necessary when a match between two structures must be found in the presence of structural noise [16, 17]. While research on graph matching is abundant, so far we know of relatively little work on dynamic graph matching.

The rest of this paper is organized as follows. First, we present the vector-space models in section 2. In section 3 we apply the

vector-space models to the task of author identification for the KDD Cup 2003 citation database and present results. In section 4, we consider approaches utilizing only the citation list of the current paper, relying on self-citations for author identification. Finally, we conclude with a discussion of the results.

2. VECTOR-SPACE METHOD

We reduce the dynamic author citation graph to a set of feature vectors for two types of *entities*: authors and papers. Each paper's feature vector represents a candidate query for identification. Each author-history feature vector summarizes an author's previous citation behavior. During the author identification process, new papers, stripped of explicitly identifiable information, are compared to labeled author-history feature vectors.

In the vector space model [18], a query is represented by a *n*-dimensional vector where *n* is the number of possible terms in the vector. Applying the vector-space model to our task, the query vector represents a new paper where the author is to be identified and the number of non-null dimensions in the vector depends on the number of (unique) citations in the paper. The query vector of a paper is compared to each author-history vector. Various weighting schemes and similarity measures could be used; we discuss these in detail below.

For dynamic graphs we need to capture the dynamics of transient relational ties. A *relational tie* [8] establishes a linkage between a pair of entities. In our example, a relational tie is established when authors cite prior papers. We give each relational tie a *weight* to indicate its strength. Each author may have multiple relationships to multiple papers. We represent each author with a vector of weights. To capture the dynamic nature of authors' publication records, our "decayed counts" technique iteratively updates each author's weight vector over time. With each new published document, the weights of old citations are reduced, so as to lessen their importance for subsequent author identification. Similar dynamic identification methods have been used to detect repetitive defaulters in large telecommunication networks [19] and in machine vision to identify visual objects in pictures with dynamic scenes [20].

2.1 Data Structure Definition

To create the paper vectors and the author-history vectors, a weight is associated with each potentially cited paper. A weight of zero is assigned when there is no relationship between two entities (e.g., an author never cited a particular paper). This structure enables us to represent an approximation of the author's entire temporal citation graph as well as new papers in the same vector space.

Definition: An entity e_i is described by a feature vector where w_{ik} is the weight assigned to the relational tie between e_i and paper p_k (1).

$$e_i = (w_{i1}, w_{i2}, \ldots, w_{il}) \quad (1)$$

2.2 Weights

The feature-vector weights can emphasize important relationships. Each individual weight is determined by some aggregation of the relationship(s) between the entity and the paper. There are different, and sometimes opposing, notions of importance that can be captured by the weights. For example, it may be important to focus on an author's current "behavior," giving higher weight to citations in more recent (historical) papers. On the other hand, it may be important to give higher weight to distinctive citations, even if they have not been cited recently. Ideally, we would like for the author-history vectors to represent frequency, recency, and idiosyncrasy of cited papers. For the vector-space model used in this paper, we consider binary weights, simple counts, and decayed counts for author-history vectors; for paper vectors, the counts are always one or zero.[1]

1) *Binary* weights are one when any relationship exists between an entity and a potentially cited paper and zero if the paper has not been cited. For example, if author A cited paper B in three papers, then the weight for the relationship corresponding to paper B in A's author-history vector is one.

2) *Simple counts* represent the total number of papers in which a citation appears. For example if author A cited paper B in three papers, then the weight corresponding to paper B in A's author-history vector is three.

3) *Decayed counts* also consider the total number of times a relationship is observed, but give more weight to the most recent citations. For this paper we use exponential smoothing: the vector for entity E_t is defined as the sum of past observations (2), where the damping factor θ determines the influence of historical values. When θ is close to one, historical values have much influence. When θ is close to zero, only the most recent observations receive non-negligible weights.

$$E_t = (1-\theta)\, e_{t-1} \oplus e_t \quad (2)$$

2.3 Matching

During the author-identification process, new papers are compared to every non-null author-history vector. Candidate match sets of author-history vectors most similar to the query are ranked and returned. For this paper, we use *cosine similarity,* which is commonly used with the vector-space model [21].

2.4 Experimental Setup

We incrementally modify the author-history vectors as new papers appear with time. When a new paper query is presented, it is compared to the contemporaneous author-history vectors. A match is considered to be correct when the method correctly identifies at least one author of the paper. We also report how often an author is in the top-10 and top-100 highest-ranked authors for the paper (of 7424 total authors).

We attempt to identify authors of papers from the KDD Cup 2003 paper archive. This database is from the Stanford Linear Accelerator Center SPIRES-HEP archive comprising High Energy Particle Physics (HEP) articles spanning the years 1992-2003. We first present results using the dynamic vector-space

[1] The paper-vector weights can be other than zero or one, as we will see for the discriminative self-citations. Using discriminative weights may improve the performance of the vector-space model as well, but here we do not consider them.

model, where author identification succeeds 26-32% of the time. Later, in Section 4, we will present self-citation-based methods that perform even better, but before doing so we will show results with the vector-space model with self-citations removed.

For the 29,514 papers, we parsed the author names from each abstract utilizing the first initial of the first name and entire last name, resulting in a total of 7424 authors. Eliminating part of the author name in some cases may contribute to a loss in data integrity for frequently occurring first-initial/last-name pairs. Name ambiguity could have opposing effects on matching accuracy. Because the total number of authors is reduced, it may serve to increase accuracy; however, it seems unlikely that an erroneous match would happen to be to someone with the same first initial and last name. It seems more likely that name ambiguity would degrade the quality of the author-history vectors, by mixing the citations of different people—making the correct author seem less similar, and thereby reducing accuracy.[2] So, using first-initial/last-name pairs seems a conservative choice.

We first sorted the papers by submission date. We incrementally updated the author-history vectors with each new paper. We only use citations to other papers within the database (*intra-database* citations).[3] At each new paper presentation, the author-history vectors comprise all past citations in the database. For estimating accuracy we used (as a "test set") the 12,387 papers submitted between 1999 and 2002.

3. VECTOR-SPACE RESULTS

Table 1a presents top-N match results for N=1,10, and 100. N=10 and N=100 are interesting because a citations-only method may be used as a preprocess to computationally expensive methods such as textual analysis. N=10 and N=100 also give a rough idea of how well similar matching methods could hope to do with the present experimental setup.

3.1 Matching accuracy

Decayed counts as weights performed best, matching 26% of the authors exactly. However, we lack information on some test papers. In particular, some papers have no (intra-database) citations. No method based solely on citations would be able to match these papers accurately to prior authors, using the current experimental setup. Once the papers containing no citations are removed, matching accuracy improves marginally (to 27%). Further investigation indicates that there are some test documents for which we are seeing the author(s) for the first time; no matching method based on prior behavior (citing behavior or otherwise) would be able to identify a new author if no author history is available. Excluding documents with no author history, matching accuracy again increases marginally (to 28%). So, although having information on both the new entity and past behavior is necessary for matching success, the lack thereof does not seem to play a major role in the observed matching accuracy. However, there are many cases where there are both intra-database citations and corresponding author history, but the vector-space methods still do not identify the author correctly. One possible explanation is that author citation behavior has changed dramatically. For example, an extreme case would be when an author moves to a completely different research topic and there is no citation overlap at all with the author's past papers. If we include only papers with at least one paper in common with at least one author, matching accuracy improves to 32%.

In sum, depending on what we are willing to exclude, the dynamic vector-space method successfully identifies correctly approximately 1/4-1/3 of the cases using only intra-database citations. Furthermore, it places an author in the top-10 67% of the time and in the top-100 89% of the time. In general, and not surprisingly, the amount of overlap between a paper's citation list and an author's history is a reasonable predictor for matching success. For example, if we consider only those papers for which an author's history has greater than 5 citations in common, we can match 40% of the time, place in the top-10 75% of the time, and place in the top-100 95% of the time.

Although identification of authors based solely on intra-database citations is a restricted setting, being able to identify authors successfully in 1/4-1/3 of the cases based only on this information (not even considering linguistic mannerisms or distinctive nuances of the research) certainly calls into question the notion that reviewing is truly double-blind.

3.2 Matching accuracy without self-citations

Self-citations occur when authors cite their own prior work. There is anecdotal evidence that self-citations are an important identifier of paper authorship, and in the following section we will show methods based only on self-citations that identify authors better than our best current vector-space method. Although a policy of eliminating self-citations is not a solution for improving double-blind review, it is interesting to investigate how much of the performance of the vector-space method is due to identifiability through self-citations. (Methods that are not dependent on counting self-citations also are of interest because author counts in a citation list are particularly easy for authors to manipulate.)

Table 1b shows the results of applying the vector-space model with decayed-count weights, using the same set-up as above except with all self-citations removed. Again, depending on what we are willing to exclude, the dynamic vector-space method successfully identifies approximately 1/6-1/5 of the cases using only intra-database citations. Furthermore, it places an author in the top-10 47% of the time and top-100 79% of the time.

Note that none of the vector-space results presented here take into account the discriminability of the citations. As we will see in the next section, weighting citations by their discriminability can improve citation-based author identification.

[2] Note that a more sophisticated system ought be able to deal with the general phenomenon of a single author with multiple, non-overlapping research interests. A corresponding modification of the current method is to cluster papers by citations and represent an author by multiple author-history vectors.

[3] A paper, of course, could cite papers not in this database. We used the information made available explicitly for the KDDCUP competition. Including extra-database citations arguably may improve matching accuracy significantly; for example, authors may be more identifiable through habits of citing particular papers from outside the immediate research community.

Weights	N	Including no cites	Including no hist	Including no overlap	At least 1 overlap
Binary	1	0.18	0.19	0.20	0.23
	10	0.48	0.50	0.53	0.60
	100	0.71	0.74	0.77	0.88
Counts	1	0.21	0.22	0.23	0.26
	10	0.49	0.52	0.54	0.62
	100	0.70	0.74	0.77	0.88
Decayed counts	1	0.26	0.27	0.28	0.32
	10	0.54	0.57	0.59	0.67
	100	0.71	0.75	0.78	0.89
Total Documents		12,387	11,777	11,262	9,869
% of test data		100%	95%	91%	80%

Table 1: ***Author identity matching accuracy for papers published during 1999-2002.*** **N is the number of top-ranked authors considered (e.g., N=1 corresponds to a correct match to the single, highest-ranked author). (a)** ***Above:*** **results including self-citations. We use three weighting schemes, binary, simple counts and decayed counts. (b)** ***Below:*** **results excluding self-citations, using decayed counts.**

The results from left to right show the matching success when we: 1) consider all test-set papers including those without any citations (no cites), 2) consider only test-set papers with at least one citation, but the author may not have been seen in the past (no hist), 3) consider only test-set papers that have at least one citation and at least one author that has authored a paper in the past—but there may be no citations in common with the current paper (no overlap), and 4) consider only test-set papers that have at least one cited paper in common with at least one of the paper's authors' histories (at least 1 overlap).

Weights	N	Including no cites	Including no hist	Including no overlap	At least 1 overlap
Decayed counts	1	0.14	0.16	0.17	0.19
	10	0.35	0.39	0.43	0.47
	100	0.60	0.67	0.73	0.79
Total Documents		12,387	11,146	10,154	8,994
% of test data		100%	90%	82%	73%

4. SELF-CITATION-BASED METHODS AND RESULTS

The vector-space methods model an author's historical citation pattern, under the assumption that citations in a new paper will follow the historical pattern. A completely different approach is to take advantage of the tendency of authors to cite their own work. This suggests a straightforward method for identifying authors based on counting cited authors: choose the author with the largest number of citations in the present paper.

Results for self-citation-based methods were generated following the same process described in detail in Section 2, except that author-history vectors are not needed. We used the same subset of papers to enable direct comparison.

Table 2 shows the accuracies of the cited-author-count method, compared with the corresponding accuracies of the vector-space model using decayed counts, for the entire test-set and for the "at least one overlap" case (discussed above). Overlap is not meaningful for the self-citation-based method, since it does not use author history; the accuracies are reported for comparison. The self-citation method clearly dominates the vector-space method, and author-identification accuracies are remarkable: almost 40% of authors can be identified exactly.

	N	Including no cites	At least one overlap
Decayed counts (vector-space model)	1	0.26	0.32
	10	0.54	0.67
	100	0.71	0.89
Cited author counts	1	0.37	0.39
	10	0.67	0.76
	100	0.78	0.88
Total Documents		12,387	9,869
% of test data		100%	80%

Table 2: Accuracy of author identification using cited-author counts. For comparison, the best results achieved from the vector-space model are listed.

Improving identification based on discriminative self-citations

Self-citation or not, a highly cited paper will be less discriminative (ceteris paribus) than a seldom-cited paper. At the extreme, if a paper is only cited by one author, it is likely to be highly discriminative for this author (in the future).

This suggests the design of methods based on discriminative citations. For example, a discriminative method could distinguish authors for (self-)citing papers that historically have seldom been cited. Specifically, rather than simply tallying the citations by author, the following discriminative self-citation methods sum up a set of paper-specific "inverse frequency" weights, which give higher weight to seldom-cited papers (cf., inverse-document frequency [21]).

In particular, we consider two variants:

1) *inverse citation-count weights* assign to citations the reciprocal of the number papers that previously have cited the paper in question.[4] For example if only papers A, B, and C cited paper D in the past, and a new query paper cites paper D, then the weight for the citation to D is 1/3. If 100 prior papers have cited the paper in question, then the weight would be 1/100.

2) *Inverse citation-frequency weights* assign to citations the log of the reciprocal of the frequency that the paper has been cited by past papers; i.e., if there are N total papers in the database and c of them cite the paper in question, the paper's weight is log(N/c).

For either weighting scheme, the score for an author for a test paper *p* is the sum of the weights of all papers by that author that are cited by *p*.

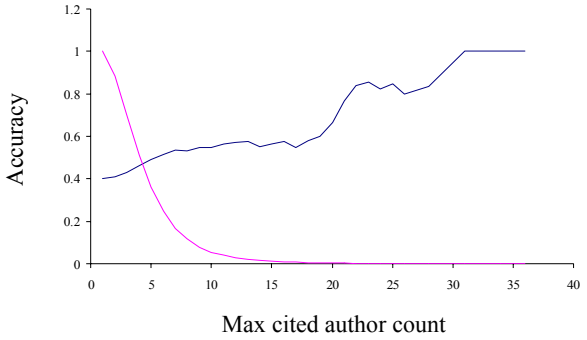

Figure 1: Accuracy as a function of the maximum number of citations to the most-cited author. The corresponding cumulative distribution of papers is also displayed (the smooth curve).

Discriminitive self-citations	N	Including no cites	At least one overlap
Inverse citation counts	1	0.38	0.43
	10	0.69	0.78
	100	0.76	0.88
Inverse citation frequencies	1	0.40	0.45
	10	0.71	0.79
	100	0.78	0.88
Total Documents		12,387	9,869
% of all data		100%	80%

Table 3: *Accuracy of author identification for papers in 1999-2002.* N is the number of top-ranked authors considered (e.g., N=1 corresponds to a correct match to the highest-ranked author). Results include self-citations.

[4]An alternative is to use the number of authors who cite the paper.

These discriminative self-citation-based methods perform remarkably well, as shown in Table 3. The inverse author frequency method performs best, identifying 40-45% of the papers' authors correctly.

All these self-citation-based methods will work only if the author appears in the citation list. However, even with this limitation, the methods have substantially higher accuracy than the vector-space methods.

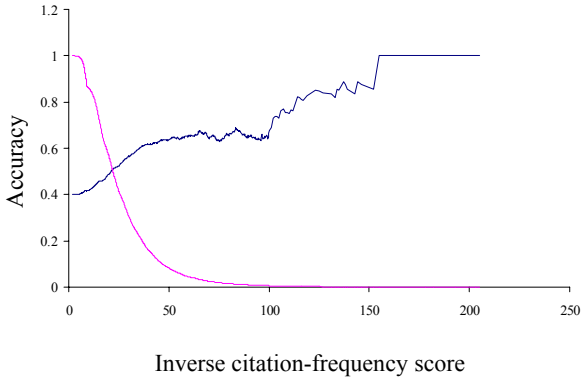

Figure 2: Accuracy of author identification as a function of matching score, using inverse citation-frequencies, along with the corresponding cumulative distribution (the smooth curve).

Figure 2 shows the accuracy of author identification as a function of the papers' maximum inverse citation-frequency scores, along with the corresponding cumulative distribution. Not surprisingly, the more citations a paper has to the most-cited author, the higher the identification accuracy (because it is more likely that this is an author of the paper).

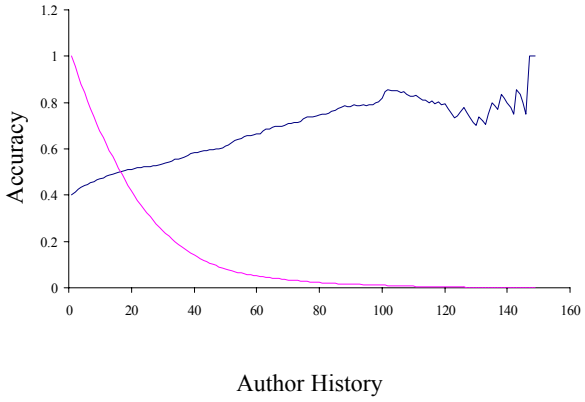

Figure 3: Accuracy of author identification for the inverse citation-frequency method as a function of the number of prior papers written by the paper's (most prolific) author, along with the corresponding cumulative distribution (the smooth curve).

Author identification accuracy increases linearly with the number of prior papers written. The more prolific authors can be identified more than half the time. The top-10% most prolific authors can be identified 60% of the time. Authors with 100 or more prior publications can be identified 85% of the time!

5. CONCLUSION

Using the KDDCUP 2003 physics-paper archive, we examine the ability to identify paper authorship automatically using only citation lists. For these papers, the discriminative self-citation-based method was able to identify authorship almost half of the time. Generally, the methods we examined were able to identify authorship between 25% and 45% of the time—based only on intra-database citations. It is possible that this may be improved by combining citation-based methods with text-based methods.

One of the main concerns surrounding the peer review process is that authors will be given preferential treatment because of their reputation. These results suggest that even when review boards institute double-blind review, many authors may be identified anyway based solely on their citations. In particular, authors with extensive publication histories can be identified well.

One limitation of this study as a criticism of double-blind review is that the results are based on published papers, not papers submitted for review. However, authors of papers submitted for review may be even easier to identify, because typically the citation list has not yet been fleshed out based on the recommendations of the reviewers, and therefore may be even more similar to an author's historical pattern of citations (and self-citations may constitute a larger fraction of the total citations). On the other hand, if an author wanted increased anonymity she could reduce the number of self-citations in the submitted version (and further obscure her publication list). This would have little effect on the main concern discussed above, because authors with good reputations would seldom be motivated to hide their identities.

The self-citation results do suggest a method for decreasing identifiability: include many citations to a well-known author.

6. ACKNOWLEDGMENTS

Thanks to Daryl Pregibon and Corinna Cortes for discussions on matching in multi-relational data mining. 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.

7. REFERENCES

[1] Meadows, A.J., *Communicating research*. 1998, San Diego: Academic Press. 266.

[2] Kassirer, J.P. and E.W. Campion, *Peer-Review - Crude and Understudied, but Indispensable.* Journal of the American Medical Association, 1994. **272**(2): p. 96-97.

[3] Hojat, M., J.S. Gonnella, and A.S. Caelleigh, *Impartial judgment by the "gatekeepers" of science: Fallibility and accountability in the peer review process.* Advances in Health Sciences Education, 2003. **8**(1): p. 75-96.

[4] Rowland, F., *The Peer Review Process.* Learned Publishing, 2002. **15**(4): p. 247-258.

[5] Williamson, A. *What happens to peer review?* In *The association of learned and professional society publishers-International Learned Journals Seminar*. 2002.

[6] Blank, R.M., *The Effects of Double-Blind versus Single-Blind Reviewing: Experimental Evidence from The American Economic Review.* The American Economic Review, 1991. **81**(5): p. 1041-1067.

[7] Ceci, S.J. and D. Peters, *How Blind Is Blind Review.* American Psychologist, 1984. **39**(12): p. 1491-1494.

[8] Wasserman, S. and K. Faust, *Social network analysis: methods and applications*. Structural analysis in the social sciences; 8. 1994, New York: Cambridge University Press. xxxi, 825.

[9] Gross, J.L. and J. Yellen, *Graph theory and its applications*. The CRC Press series on discrete mathematics and its applications. 1999, Boca Raton, Fla.: CRC Press. 585.

[10] Donohue, J.C., *Understanding scientific literatures: a bibliometric approach*. 1974, Cambridge,: MIT Press. xiii, 101.

[11] Burt, R.S., *Applied Network Analysis: A Methodological Introduction*. 1983, Beverly Hills, CA: Sage Publications. 262-282.

[12] Egghe, L. and R. Rousseau, *Co-citation, bibliographic coupling and a characterization of lattice citation networks.* Scientometrics, 2002. **55**(3): p. 349-361.

[13] de Vel, O., et al., *Mining e-mail content for author identification forensics.* Special Interest Group on Management of Data Record, 2001. **30**(4): p. 55-64.

[14] Diederich, J., et al., *Authorship attribution with support vector machines.* Applied Intelligence, 2003. **19**(1-2): p. 109-123.

[15] Corney, M., et al. *Gender-Preferential Text Mining of E-mail Discourse*. In *18th Annual Computer Security Applications Conference*. 2002. San Diego California.

[16] Bengoetxea, E., et al., *Inexact graph matching by means of estimation of distribution algorithms.* Pattern Recognition, 2002. **35**(12): p. 2867-2880.

[17] Medasani, S., R. Krishnapuram, and Y. Choi, *Graph matching by relaxation of fuzzy assignments.* IEEE Transactions on Fuzzy Systems, 2001. **9**(1): p. 173-182.

[18] Raghavan, V.V. and S.K.M. Wong, *A critical analysis of vector space model for information retrieval.* Journal of the American Society for Information Science, 1986. **37**(5): p. 279--287.

[19] Cortes, C., D. Pregibon, and C.T. and Volinsky. *Communities of Interest for Dynamic Graphs*. *In the Proceedings of Knowlege Discovery and Data Mining Conference*. 2002. Edmonton, Canada.

[20] Wilson, R.C. and E.R. Hancock, *Relational Matching with Dynamic Graph Structures.* Proceedings of the Fifth International Conference on Computer Vision, 1995: p. 450-456.

[21] Baeza-Yates, R. and B.d.A.N. Ribeiro, *Modern information retrieval*. 1999, New York

---

About the authors:

Shawndra Hill is a doctoral candidate in the Information Systems Department at New York University's Stern School of Business.

Foster Provost is Associate Professor in the Information, Operations, and Management Sciences Department at New York University's Stern School of Business