Citations Beyond Self Citations: Identifying Authors, Affiliations, and Nationalities in Scientific Papers

Yoshitomo Matsubara
University of California, Irvine
yoshitom@uci.edu

Sameer Singh
University of California, Irvine
sameer@uci.edu

Abstract

The question of the utility of the blind peer-review system is fundamental to scientific research. Some studies investigate exactly how "blind" the papers are in the double-blind review system by manually or automatically identifying the true authors, mainly suggesting the number of self-citations in the submitted manuscripts as the primary signal for identity. However, related studies on the automated approaches are limited by the sizes of their datasets and the restricted experimental setup, thus they lack practical insights into the blind review process. Using the large Microsoft Academic Graph, we train models that identify authors, affiliations, and nationalities of the affiliations for anonymous papers, with 40.3%, 47.9% and 86.0% accuracy respectively from the top-10 guesses. Further analysis on the results leads to interesting findings *e.g.*, 93.8% of test papers written by Microsoft are identified with top-10 guesses. The experimental results show, against conventional belief, that the self-citations are no more informative than looking at the common citations, thus suggesting that removing self-citations is not sufficient for authors to maintain their anonymity.

1 Introduction

Scientific publications play an important role in dissemination of advances, and they are often reviewed and accepted by professionals in the domain before publication to maintain quality. In order to avoid unfairness due to identity, affiliation, and nationality biases, peer review systems have been studied extensively (Yankauer, 1991; Blank, 1991; Lee et al., 2013), including analysis of the opinions of venue editors (Brown, 2007; Baggs et al., 2008) and evaluation of review systems (Yankauer, 1991; Tomkins et al., 2017). It is widely believed that a possible solution for avoiding biases is to keep the author identity blind to the reviewers, called double-blind review, as opposed to only hiding the identity of the reviewers, as in single-blind review (Lee et al., 2013). Since some personal information (*e.g.*, author, affiliation and nationality) could implicitly affect the review results (Lee et al., 2013), these procedures are required to keep them anonymous in double-blind review, but this is not foolproof. For example, experienced reviewers could identify some of the authors in a submitted manuscript from the context. In addition, the citation list in the submitted manuscript can be useful in identifying them (Brown, 2007), but is indispensable as it plays an important role in the reviewing process to refer readers to related work and emphasize how the manuscript differs from the cited work.

To investigate blindness in double-blind review systems, Hill and Provost (2003) and Payer et al. (2015) train a classifier to predict the authors, and analyze the results. However, they focus primarily on the utility of self-citations in the submitted manuscripts as a key to identification (Mahoney et al., 1978; Yankauer, 1991; Hill and Provost, 2003; Payer et al., 2015), and do not take author's citation history beyond just self-citations into account. The experiment design in these studies is also limited: they use relatively small datasets, include papers only from a specific domain (*e.g.*, physics (Hill and Provost, 2003), computer science (Payer et al., 2015) or natural language processing (Caragea et al., 2019)), and pre-select the set of papers and authors for evaluation (Payer et al., 2015; Caragea et al., 2019). Furthermore, they focus on author identification, whereas knowing affiliation and the nationality also introduces biases in the reviewing process (Lee et al., 2013).

In this paper, we use the task of author identity, affiliation, and nationality predictions to analyze the extent to which citation patterns matter, evaluate our approach on large-scale datasets in many domains, and provide detailed insights into

the ways in which identity is leaked. We describe the following contributions:

1. We propose approaches to identify the aspects of the citation patterns that enable us to guess the authors, affiliations, and nationalities accurately. To the best of our knowledge, this is the first study to do so. Though related studies mainly suggest authors avoid self-citations for increasing anonymity of submitted papers, we show that overlap between the citations in the paper and the author's previous citations is an incredibly strong signal, even stronger than self-citations in some settings.
2. Our empirical study is performed on (i) a real-world large-scale dataset with various fields of study (computer science, engineering, mathematics, and social science), (ii) study different relations between papers and authors, and (iii) two identification situations: "guess-at-least-one" and "cold start". For the former, we identify authors, affiliations and nationalities of the affiliations with 40.3%, 47.9% and 86.0% accuracy respectively, from the top-10 guesses. For the latter, we focus on papers whose authors are not "*guessable*", and find that the nationalities are still identifiable.
3. We perform further analysis on the results to answer some common questions on blind-review systems: "*Which authors are most identifiable in a paper?*", "*Are prominent affiliations easier to identify?*", and "*Are double-blind reviewed papers more anonymized than single-blind?*". One of the interesting findings is that 93.8% of test papers written by a prominent company can be identified with top-10 guesses.

The dataset used in this work is publicly available, and the complete source code for processing the data and running the experiments is also available.[2]

2 Related work

Here, we summarize related work, and describe their limitations in analyzing anonymity in the blind review systems.

2.1 Citation Analysis and Application

There are several studies that propose applications using citation networks (Dong et al., 2017), and they are not limited to applications of scientific papers in academia. Fu et al. (2015, 2016) study patent citation recommendation and propose a citation network modeling. Levin et al. (2013) introduce new features for citation-network-based similarity metric and feature conjunctions for author disambiguation, and it outperforms the clustering with features from prior work. Fister et al. (2016) define citation cartel as a problem arising in scientific publishing, and they introduce an algorithm to discover the cartels in citation networks using a multi-layer network. Petersen et al. (2010) propose the methods for measuring the citation and productivity of scientists, and examine the cumulative citation statistics of individual authors by leveraging six different journal paper datasets. Though a study of Su et al. (2017) is not a citation related work, it proposes an approach to de-anonymize web browsing histories with social networks and link them to social media profiles. Kang et al. (2018) publish the first dataset of scientific peer reviews, including drafts and the decisions in ACL, CoNLL, NeurIPS and ICLR. Using the published dataset, they also present simple models to predict the accept/reject decisions and numerical scores of review aspects.

2.2 Blind Review and Author Identification

Blind review systems in conferences and journals have been addressed for decades, and have attracted researchers' attention recently (Blank, 1991; Brown, 2007; Lee et al., 2013). For instance, Snodgrass (2006) summarizes previous studies of the various aspects in blind reviewing within a large number of disciplines, and discusses the efficacy of blinding while mentioning how blind submitted/published papers are in different studies. Tomkins et al. (2017) show an example of affiliation bias in the reviewing process. They performed an experiment in the reviewing process of WSDM 2017, which considers the behavior of the program committee (PC) members only, and the members are randomly split into two groups of equal size: single-blind and double-blind PCs. They report that single-blind reviewers bid for 22% more papers, and preferentially bid for papers from top institutions. Bharadhwaj et al. (2020) discuss the relation between de-anonymization of authors through arXiv preprints and acceptance of a research paper at a (nominally) double-blind venue. Specifically, they create a dataset of ICLR 2020 and 2019 submissions, and present key inferences obtained by analyzing the dataset such as "releasing preprints on arXiv has a positive correlation with

[2] https://github.com/yoshitomo-matsubara/guess-blind-entities

acceptance rates of papers by well known authors."

Some studies attempt to manually identify authors and affiliations in submitted manuscripts. Yankauer (1991) sent a short questionnaire the reviewers of American Journal of Public Health for asking them to identify the author and/or institution of submitted manuscripts, and reported that blinding could be considered successful 53% of time. Justice et al. (1998) examine whether masking reviewers to author identity improves the peer review quality. Through a controlled trial for external reviews of manuscripts submitted to five different journals, they conclude that masking fails to the identity of well known authors, and may not improve the fairness of review.

In addition to the manual identification studies, some researchers propose automatic approaches to guess authors in published papers. Table 1 summarizes datasets in other studies. To the best of our knowledge, Hill and Provost (2003) first propose automatic methods using citation information for author identification and perform an experiment with a dataset, that consists of physics papers in the arXiv High Energy Particle Physics between 1992 and 2003. Payer et al. (2015) propose deAnon, a multimodal approach to deanonymize authors of academic papers. They perform experiments with papers in the proceedings of 17 different computer science related conferences from 1996 to 2012. Similarly, Caragea et al. (2019) address a similar research question, and train convolutional neural networks on the datasets of the prefiltered ACL and EMNLP papers, using various types of features such as context, style, and reference.

However, there are some biased observations in their work. As shown in Table 1, one of the biggest concerns lies in their datasets. They use only one major field dataset in their work: physics (Hill and Provost, 2003) , computer science (Payer et al., 2015) and natural language processing (Caragea et al., 2019), but it would be not enough to discuss if their approaches actually work in various fields of study. The second biggest concern is that they understate a possibility that there are also papers where no authors can be found in the training dataset (Payer et al., 2015; Caragea et al., 2019). Especially in Payer et al. (2015)'s work, the authors do not mention the possibility, but achieve 100% accuracy after trying all guesses for each paper in their *guess-one*, *guess-most-productive-one* and *guess-all* scenarios even though it is very difficult in general to find papers where all the authors are seen in the training dataset.

Furthermore, they focus only on productive authors who have at least three papers in the training dataset, and the numbers of candidates in training and test papers can be considered very limited. Similarly, Caragea et al. (2019) exclude any authors with less than three papers from their datasets after an author name normalization process described in Section 4.3. Hill and Provost (2003) argue that there are some test papers for which they did not see the author(s) in their training dataset. However, the lack of true authors' citation histories does not seem to strongly affect their observed matching accuracy, and it can be caused by the scale of the dataset. Also, their studies do not cover either affiliation or nationality (including cold start scenario), which could cause affiliation and nationality biases (Lee et al., 2013) if they are identifiable.

3 Identification Approach

Training and test datasets are independently prepared, and papers in the training dataset are older than those in the test dataset. We extract features from the training dataset to model each author's citation pattern, and the entity also can be affiliation or nationality depending on what we guess in the test papers. Building entity models, we score each entity based on its extracted features for a test paper, and sort the scores for the paper to rank all the entities. We describe the detail of each process in the following sections.

3.1 Citation Features

Scientific papers have references to introduce related work to readers and sometimes compare the results with the work in order to emphasize the difference between them. We assume that authors have their own citation patterns, and it can be a clue to guess authors in papers. They would repeatedly cite the same papers and their own publication if the projects and fields are similar to their previous ones. Also, we assume that the citation list in a paper would not dramatically change between before and after the blind-review process, since we are limited in access to the published papers only.

In addition to citation features (Hill and Provost, 2003), Payer et al. (2015) and Caragea et al. (2019) use contextual features. As discussed in (Narayanan et al., 2012; Rosen-Zvi et al., 2004), author-topic model and writing style would be hints

Table 1: Dataset comparison with other studies.

	Hill and Provost (2003)	Payer et al. (2015)	Caragea et al. (2019)	Our work
Domains	Physics	CS	NLP	All, CS, Eng., Math, Soc. Sci.
#authors	7,424	1,405	262 & 922	22k - 2M
#papers	29,514	3,894	622 & 3,011	231k - 825k

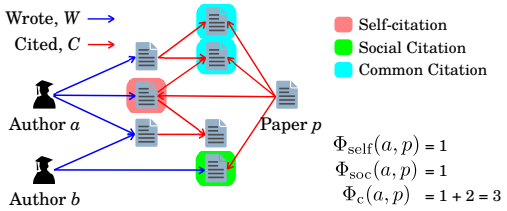

Figure 1: Example of self-, social and common citations $\Phi_{\{\text{self, soc, c}\}}(a, p)$ for author a and paper p.

to identify authors. In this work, however, we only use citation and publication histories for identification. This also reduces computational load in training and test processes and enables us to further analyze the performances in various situations focused on citation features. In the following approaches, the models skip scoring candidate authors (entities) given a test paper if they have no citation features (all zero(s)) since this work focuses on citation pattern in the identification problems.

Figure 1 illustrates an example citation graph with red and blue edges from $x \rightarrow y$ indicating x cited y and x wrote y, respectively. We focus here on three types of citations described in the following sections: self, social, and common citations.

3.2 Self-citations, SC

As discussed in these studies (Mahoney et al., 1978; Yankauer, 1991; Hill and Provost, 2003; Payer et al., 2015), self-citations can be a clue in identification. The Self-citation (SC) model calculates how many papers written by author a are cited by paper p based on his/her publication history

$$\Phi_{\text{self}}(a, p) = \sum_{r \in \text{Ref}_p} W(a, r)\,,$$

$$W(a, p) = \begin{cases} 1 & \text{if } a \text{ wrote } p \\ 0 & \text{otherwise} \end{cases},$$

where p is a blind (test) paper, and a is a candidate author seen in the training dataset. Ref_p is the set of paper IDs cited by paper p. In Figure 1, a wrote three different papers, and one of them is cited by p i.e., $(\Phi_{\text{self}}(a, p) = 1)$, assuming a wrote p.

Hill and Provost (2003) use inverse citation-frequency (icf) for weighted scoring for self-citations to incorporate importance of the self-citation. We include this in our SC model as well:

$$\Phi^{\text{icf}}_{\text{self}}(a, p) = \sum_{r \in \text{Ref}_p} W(a, r) \cdot icf(r) \qquad (1)$$

$$icf(r) = \log\Big(\frac{N_{\text{tr}}}{1 + \sum_{p' \in P_*} C(p', r)}\Big)\,,$$

$$C(p, r) = \begin{cases} 1 & \text{if } p \text{ cited } r \\ 0 & \text{otherwise} \end{cases},$$

where P_* denotes the set of papers in the training dataset, $N_{\text{tr}} = |P_*|$ is the number of papers, and A is the set of all authors in the training dataset.

3.3 Social citations, SocC

Instead of self-citations, it is also common to cite papers written by past collaborators. In this work, we call such citations social citations. Though this model itself will not be as powerful as the SC model, the social citation feature helps us identify potential connections between a test paper and candidates (authors) as this approach covers the publication histories of the past collaborators given an author. Social citation score is defined as:

$$\Phi_{\text{soc}}(a, p) = \sum_{r \in \text{Ref}_p} \sum_{a_c \in A_a} W(a_c, r), \qquad (2)$$

where A_a is the set of authors who wrote a paper with author a. In Figure 1, author a wrote a paper with author b, and p cited a paper written by b. Then, the social citation count is one.

Similar to the SC model, our SocC model uses the weighted score:

$$\Phi^{\text{icf}}_{\text{soc}}(a, p) = \sum_{r \in \text{Ref}_p} \sum_{a_c \in A_a} W(a_c, r) \cdot icf(r). \qquad (3)$$

3.4 Common Citations, CC

Apart from self and social citations, another clue to the identity might be in all past citations (even ones that are not self or social). Common Citation (CC)

Table 2: Features used for our combined model.

Feature Name	Feature Value
Average icf-weighted CC score	$\frac{\Phi_{\mathrm{c}}^{\mathrm{icf}}(a,p)}{\lvert\mathrm{Ref}_p\rvert}$
CC coverage	$\frac{\lvert\mathrm{Ref}_p \wedge \mathrm{Ref}_a^*\rvert}{\lvert\mathrm{Ref}_p\rvert}$
Average SocC score	$\frac{\Phi_{\mathrm{soc}}^{\mathrm{icf}}(a,p)}{\lvert\mathrm{Ref}_p\rvert}$
SocC coverage	$\frac{\lvert\mathrm{Ref}_p \wedge \mathrm{Pub}_{A_a}\rvert}{\lvert\mathrm{Ref}_p\rvert}$
icf-weighted SC score	$\Phi_{\mathrm{self}}^{\mathrm{icf}}(a,p)$
SC score	$\Phi_{\mathrm{self}}(a,p)$

Ref_a^*: set of paper IDs cited by papers written by a in the training dataset, while Pub_{A_a}: set of papers written by past collaborators of author a.

model thus calculates how many times in author a cites each of the papers cited by paper p:

$$\Phi_c(a,p) = \sum_{r \in \mathrm{Ref}_p} \sum_{p'_a \in P_a} C(p'_a, r)\,, \qquad (4)$$

where P_a is the set of a's papers in the training dataset. In Figure 1, the paper p cites two of the papers cited by a, and the author's common citation count is three. We also include a weighted version:

$$\Phi_{\mathrm{c}}^{\mathrm{icf}}(a,p) = \sum_{r \in \mathrm{Ref}_p} \sum_{p'_a \in P_a} C(p'_a, r) \cdot icf(r). \qquad (5)$$

3.5 Learning a Classifier

In addition to separately using the SC, SocC and CC models, we introduce a combined model (Full) that uses all the citation features. We estimate the parameters of features by the mini-batch gradient descent method. Due the cost of computing softmax function over all possible authors for a paper, we use negative sampling, similar to (Mikolov et al., 2013), leading to the following loss:

$$l(\{a_i, p_i\}, \boldsymbol{\theta}) = \frac{1}{K} \sum_{i=1}^{K} \Big(\log \sigma\big(\boldsymbol{\theta} \cdot \boldsymbol{\phi}(a_i, p_i)\big) - \frac{1}{|\bar{A}_{p_i}|} \sum_{\bar{a} \in \bar{A}_{p_i}} \log \sigma\big(\boldsymbol{\theta} \cdot \boldsymbol{\phi}(\bar{a}, p_i)\big) \Big) - \lambda ||\boldsymbol{\theta}||_2^2 \qquad (6)$$

where $\{a_i, p_i\}$ is a set of pairs of authors and their papers, and $\boldsymbol{\theta}$ is 7-dimensional estimated parameter vector. $\boldsymbol{\phi}(a_i, p_i)$ contains a bias term and features shown in Table 2, and K is the batch size. $\bar{A}_{p_i}$ is a set of randomly sampled authors as negative samples given paper p_i, and λ is a hyperparameter for regularization. Note that these parameters $\boldsymbol{\theta}$ are shared across all the authors in the dataset.

4 Experimental Setup

We define some terms and variables used in the following sections, and then describe the MAG dataset and how we develop benchmarks from it.

4.1 Evaluation Setup

We consider three different entity disambiguation scenarios: author, affiliation, and nationality. For each, our primary evaluation metric is *hits at least*, **HAL**M@k, accuracy of our guesses. If our top-k ranking hits at least M of all the true entities in a test paper, it is considered successfully guessed. M is typically fixed at 1 in the related studies (Blank, 1991; Yankauer, 1991; Justice et al., 1998; Hill and Provost, 2003; Payer et al., 2015; Caragea et al., 2019). Similarly, the range of k is 1-100 (Hill and Provost, 2003), 1-1000 (Payer et al., 2015) and 10 (Caragea et al., 2019) in the previous work respectively. We also consider an evaluation where we set k to X, the number of the *true* entities of a test paper (*i.e.*, each test paper has a different X.

Additionally, we differentiate between *guessable* and *not guessable* papers. We call a test paper ***guessable*** if at least M of all the true entities in the training set have any (non-zero) citation feature used in a model. If M is greater than the number of the true entities in a test paper, it is not ***guessable***.

4.2 Dataset: Microsoft Academic Graph

The Microsoft Academic Graph (MAG) is a large heterogeneous graph of academic entities provided by Microsoft. For paper and author entities, Sinha et al. (2015) collect data from publisher feeds (*e.g.*, IEEE and ACM) and web-page indexed by Bing. They also report that often the quality of the feeds from publishers are significantly better, although the majority of their data come from the indexed pages. The MAG was used in the KDD Cup 2016 for measuring the impact of research institutions and in the WSDM Cup 2016 for entity ranking challenge. The MAG is much larger and more diverse than datasets used in related studies (Hill and Provost, 2003; Payer et al., 2015; Caragea et al., 2019), and uses disambiguated entity IDs. Since some authors seem to be assigned to different author IDs though they look identical, we perform author disambiguation in a more conservative method (Section 4.3) than those in the previous work (Hill and Provost, 2003; Caragea et al., 2019). We use the dataset released in February 2016, thus it includes very few papers published in 2016 than in

the years earlier. Some entries do not have all the attributes we need; we discard such entries.

4.3 Author Disambiguation

It would be ideal if an author name uniquely identifies the entity. In practice, however, an author name tends to be directed to different entities, and an entity may correspond to multiple names (*e.g.*, misspelling and shortened names). Hill and Provost (2003) used the dataset[3] released for KDD Cup 2003. Since this dataset does not contain author IDs, they performed author name disambiguation on the dataset by using author's initial of the first name and entire last name, and Caragea et al. (2019) used the same technique.

Though Hill and Provost (2003) consider the method conservative, it seems rather rough when we tried to reproduce the result. We found that there are 12,625 unique author names, and their disambiguation method resulted in 8,625 unique shortened author names. However, 883 of them have potential name conflicts. Taking an example from the result, "Tadaoki Uesugi" and "Tomoko Uesugi" are considered identical as "T Uesugi", but their names look completely different. Another example is with shortened name; there is a conflict between "A Suzuki", "Alfredo Suzuki" and "Akira Suzuki" though it would make sense if there were only one pair of "A Suzuki" and "Alfredo Suzuki" (or "Akira Suzuki") in the dataset.

The MAG dataset contains author IDs, but there still remains some ambiguity of authors. One of the possible reasons is that some authors may have moved to different affiliations and their new author IDs were generated. Leveraging some of the knowledge in KDD Cup 2013 (author disambiguation challenge) (Chin et al., 2013), we merge authors into one entity if and only if they meet all the following conditions: (1) they have identical full names, and (2) have at least one common past collaborator. This policy reduces the number of unique author IDs in our extracted datasets by about 4%. It may be still incomplete, but it is more conservative and would bias our results less than related work (Hill and Provost, 2003; Caragea et al., 2019).

4.4 Extracted Datasets

Since the MAG dataset is significantly larger than the datasets used in the previous studies (Hill and Provost, 2003; Payer et al., 2015; Caragea et al., 2019), we extract five different datasets from the MAG dataset: randomly sampled, computer science, engineering, mathematics, and social science datasets. All these datasets consist of papers published between 2010 and 2016, and we split the datasets into training (from 2010 to 2014) and test (from 2015 to 2016) datasets. As we mentioned in Section 4.2, the original dataset includes few papers published in 2016 due to its release date. Note that the test datasets include over 20% of the test papers all of whose authors are not found in the training datasets since these training and test datasets are independently prepared.

The first dataset (MAG(10%)) is composed of randomly sampled papers to extract 10% of the whole dataset, and it is most diverse with respect to fields of study among the five datasets. All the other datasets are extracted based on the venue list for each field. For efficiency, it is reasonable to filter candidates (and papers in training dataset) by their fields given a paper because reviewers will know the fields of their venues. Here, an extracted candidate has at least one paper published at a venue in the field defined below, and papers in the training dataset consists of papers written by extracted candidates. Though some papers may not be guessable because of the filter, we consider the possibility to keep our experimental design unbiased (*i.e.*, we do not discard test papers responding to the filtered training dataset). For computer science (CS), we extract papers presented at any of the 60 different venues in a list based on CSRankings[4]. We also create lists of conferences based on Scimago Journal & Country Rank[5] for engineering (Eng.), mathematics (Math), and social science (Soc. Sci.), and the lists consist of 60, 60, and 34 venues respectively. Table 3 shows the statistics of each dataset in author identification. Because of few venues of social science in the original dataset, the dataset is smaller than the others, but still larger than those used in the previous studies (Hill and Provost, 2003; Payer et al., 2015; Caragea et al., 2019).

4.5 Entity Conversion

We also use the above datasets for affiliation and nationality identifications (see Tables 4 and 5 for details). Since some papers in the datasets lack affiliation information, we drop papers from the

[3] https://www.cs.cornell.edu/projects/kddcup/datasets.html

[4] http://csrankings.org/

[5] http://www.scimagojr.com/

Table 3: Author Identification: Statistics of training (2010-2014) and test (2015-2016) datasets.

Dataset	Avg. X	# author IDs		# unique papers	
	test	training	test	training	test (guessable)
MAG(10%)	4.97	2,138,060	484,215	715,968	110,565 (34.1%)
CS	3.81	61,621	19,284	449,875	6,363 (64.7%)
Eng.	3.77	45,731	18,537	391,768	6,065 (48.0%)
Math	3.29	29,950	4,957	269,015	1,723 (53.6%)
Soc. Sci.	3.12	22,059	1,737	231,110	603 (28.7%)

Table 4: Affiliation Identification: Statistics of training (2010-2014) and test (2015-2016) datasets.

Dataset	Avg. X	# affiliation IDs		# unique papers	
	test	training	test	training	test (guessable)
MAG(10%)	1.72	12,416	6,441	289,748	34,927 (78.0%)
CS	1.62	8,487	1,506	260,990	5,738 (93.0%)
Eng.	1.50	8,043	1,646	222,229	5,386 (88.6%)
Math	1.51	7,124	698	153,629	1,265 (94.3%)
Soc. Sci.	1.43	6,597	401	128,718	432 (79.8%)

Table 5: Nationality Identification: Statistics of training (2010-2014) and test (2015-2016) datasets.

Dataset	Avg. X	# nationality IDs		# unique papers	
	test	training	test	training	test (guessable)
MAG(10%)	1.16	130	112	190,026	23,579 (75.5%)
CS	1.16	115	64	194,378	4,073 (89.7%)
Eng.	1.17	108	62	168,631	3,738 (83.9%)
Math	1.16	108	49	114,854	895 (91.8%)
Soc. Sci.	1.08	106	34	98,665	322 (73.6%)

training and test datasets used in affiliation identification if we cannot find at least one affiliation in each of the papers. Since the original dataset does not have nationality information for each affiliation, we perform substring matching for affiliation name based on the information by LinkedIn[6] and Webometrics[7] in order to convert an affiliation to its nationality. Similarly, we drop papers from nationality identification if we cannot find at least one nationality in each of the papers. Note that industrial affiliations may have their offices at several countries, and therefore it is difficult to use their names when converting an affiliation to its nationality. For this reason, we use academic affiliations only in affiliation identification.

Basically, each reference paper can be cited by several published papers, and similarly each published paper can be written by several authors. In contrast, each author (ID) belongs to an affiliation (ID), and an academic affiliation is in a nationality. For this dataset, we can also say that the nationality-affiliation and affiliation-author relationships are single-to-single, and the author-published paper and published paper-reference paper relationships are single-to-many. Authorship and citations of an affiliation are the total papers/citations of their authors, respectively, and similarly for authorship/citations of a nationality.

[6] https://www.linkedin.com/
[7] http://www.webometrics.info/

4.6 Baseline approaches

We extract several sub-datasets based on fields of study from the original dataset. Since the scale of the dataset depends on the field, we use a random scoring approach (Rand) as a baseline to relatively evaluate performance for each dataset. The score is randomly generated between 0 and 1. We also use another random scoring approach (Rand(S)) that skips scoring the candidate authors in a test paper if their citation histories do not include any of the papers cited by the test paper. Since the SC model is based on Hill and Provost (2003), it is also a baseline approach.

5 Experiments and Results

Using various approaches explained above, we perform experiments in two different identification scenarios: "guess-at-least-on" and "cold start". Through the first experiment, we show how anonymized a paper is in each of author, affiliation and nationality identifications. In the second experiment, we show that there remain identity leaks even when no authors in a paper are identifiable.

5.1 *Guess-At-Least-One* Identification

In this experiment, we aim to guess at least one author / affiliation / nationality ($M = 1$), and evaluate HAL1 performances of the five different approaches. If our top k ranking (guesses) includes at least one author in a given paper, the guess is considered successful. Obviously, a paper is less anonymous if we can identify at least one entity (author / affiliation / nationality) in the paper with few guesses. Tables 6-10 show identification performances with five different datasets. The average of Xs and the percentage of the guessable papers in each dataset are given in Tables 3-5.

Overall, our combined model consistently achieves the best performances in the author identification with the datasets, and in the affiliation and nationality identifications the performances of the common citation approach are comparable to those of our combined model. As for the social citation approach, interestingly, it performs better in author identification than in affiliation and nationality identifications though all the other approaches

Table 6: Guess-At-Least-One Scenario: Identification performances with randomly sampled dataset.

MAG(10%)	**Author Identification [%]**				**Affiliation Identification [%]**				**Nationality Identification [%]**		
Top	X	10	100	1000	X	10	100	1000	X	10	100
Rand	0.003	0.0009	0.01	0.089	0.028	0.123	1.37	12.7	1.10	8.60	79.7
Rand(S)	1.63	2.67	12.5	27.8	2.66	9.20	31.2	42.8	11.3	52.8	75.5
SC	8.33	9.71	10.8	10.8	5.67	7.25	7.34	7.34	11.1	12.9	12.9
SocC	6.95	8.62	11.3	11.7	0.544	1.60	7.76	18.7	0.674	3.72	16.5
CC	12.4	15.4	25.5	31.7	11.5	22.9	38.6	42.9	37.3	71.1	75.5
Full	**13.4**	**16.5**	**26.8**	**32.9**	**12.0**	**23.6**	**40.1**	**48.8**	**37.6**	**71.7**	**77.9**

Table 7: Guess-At-Least-One Scenario: Identification performances with computer science dataset.

CS	**Author Identification [%]**				**Affiliation Identification [%]**				**Nationality Identification [%]**		
Top	X	10	100	1000	X	10	100	1000	X	10	100
Rand	0.00	0.015	0.283	2.81	0.00	0.157	2.04	18.1	1.74	10.5	88.8
Rand(S)	2.40	5.30	25.7	55.1	2.09	9.22	46.2	74.2	9.18	51.1	89.7
SC	27.0	34.2	38.1	38.1	**23.4**	37.4	38.3	38.3	45.1	56.6	56.6
SocC	15.5	23.3	38.6	43.5	1.17	4.98	30.1	68.3	1.17	6.41	59.4
CC	24.6	33.9	52.3	60.5	20.1	43.7	69.2	74.2	**54.1**	85.1	89.7
Full	**30.3**	**40.3**	**56.4**	**63.9**	22.7	**47.9**	**71.3**	**79.9**	43.1	**86.0**	**93.0**

Table 8: Guess-At-Least-One Scenario: Identification performances with engineering dataset.

Eng.	**Author Identification [%]**				**Affiliation Identification [%]**				**Nationality Identification [%]**		
Top	X	10	100	1000	X	10	100	1000	X	10	100
Rand	0.00	0.033	0.313	3.13	0.037	0.149	2.01	17.7	1.39	11.7	93.5
Rand(S)	3.15	6.43	25.6	44.3	2.64	10.4	42.8	58.9	10.2	53.6	83.9
SC	19.0	22.1	22.9	22.9	15.0	21.1	21.2	21.2	31.4	37.1	37.1
SocC	9.73	14.5	22.4	23.5	0.613	2.73	17.8	45.7	0.00	1.55	25.2
CC	18.9	25.3	39.5	44.9	15.4	32.1	53.8	58.9	**44.4**	78.3	83.9
Full	**22.4**	**29.8**	**42.4**	**47.7**	**16.7**	**34.6**	**56.2**	**66.4**	40.5	**79.5**	**88.6**

Table 9: Guess-At-Least-One Scenario: Identification performances with mathematics dataset.

Math	**Author Identification [%]**				**Affiliation Identification [%]**				**Nationality Identification [%]**		
Top	X	10	100	1000	X	10	100	1000	X	10	100
Rand	0.00	0.058	0.464	3.31	0.00	0.158	2.37	21.0	1.34	9.60	94.0
Rand(S)	3.83	7.66	31.5	51.2	2.92	13.2	51.9	70.4	10.8	59.7	91.8
SC	23.7	27.0	27.3	27.3	21.6	28.2	28.3	28.3	35.9	43.6	43.6
SocC	11.7	19.4	26.5	27.3	0.395	2.92	19.7	48.3	0.670	5.47	42.7
CC	22.1	32.8	46.5	51.2	20.6	43.1	67.2	70.4	**50.7**	87.0	91.8
Full	**26.5**	**36.3**	**49.4**	**53.6**	**22.5**	**46.0**	**69.7**	**78.7**	47.6	**87.6**	**94.3**

Table 10: Guess-At-Least-One Scenario: Identification performances with social science dataset.

Soc. Sci.	**Author Identification [%]**				**Affiliation Identification [%]**				**Nationality Identification [%]**		
Top	X	10	100	1000	X	10	100	1000	X	10	100
Rand	0.00	0.00	0.166	2.32	0.00	0.00	2.78	19.2	2.17	9.94	97.2
Rand(S)	3.65	6.14	18.2	26.5	3.94	8.30	27.8	38.2	15.8	53.7	73.6
SC	14.1	16.6	17.1	17.1	**14.8**	19.4	19.4	19.4	34.8	36.6	36.6
SocC	7.13	9.95	15.4	15.8	1.85	4.40	13.7	32.9	0.00	1.55	25.2
CC	12.8	17.9	24.2	26.9	11.1	24.8	35.9	38.2	**51.2**	69.9	73.6
Full	**15.4**	**21.1**	**26.7**	**28.7**	13.4	**27.8**	**39.4**	**46.5**	**51.2**	**71.1**	**79.8**

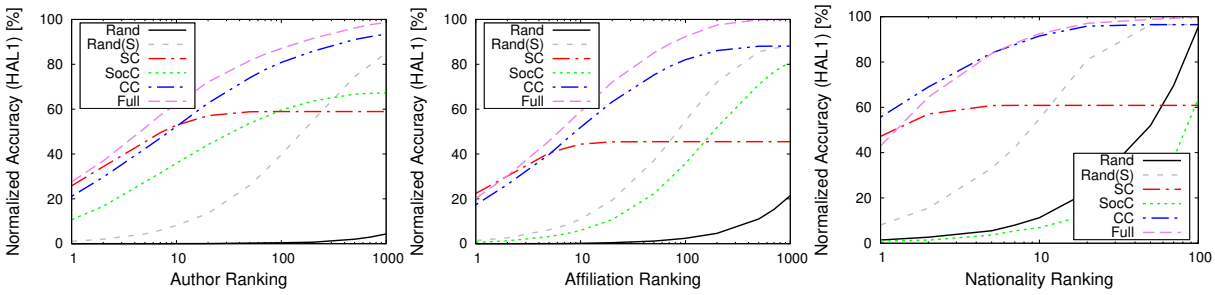

Figure 2: **Author (a), Affiliation (b) and Nationality (c) Identifications:** Normalized performances (divided by the percentage of guessable papers = 64.7, 84.3, 89.7[%] respectively) of five different approaches with CS dataset.

perform best in nationality identification. In addition, as we expected, filtering training datasets (candidates) by venues (fields of study) is effective to guess blind entities in papers of the fields though it is more difficult to guess entities in papers of the randomly sampled and social science datasets because of their smaller percentages of the guessable papers in the datasets.

Figure 2 illustrate the relations between rankings and normalized accuracies with the computer science dataset in author, affiliation and nationality identifications. The self-citation performances converge faster than other approaches using common citation, and this implies that test papers are more likely to have common citations than self-citations. In addition, the performance difference between the SC and our CC (and combined) models are significantly increasing after top 10 choices. Compared to author and affiliation identifications, the number of candidate countries in nationality identification is much smaller, and it could help us easily guess nationalities in test papers.

Some previous studies (Mahoney et al., 1978; Yankauer, 1991; Hill and Provost, 2003; Payer et al., 2015) argue that citing their own papers can be a clue to guess them in their submitted manuscript, and Hill and Provost (2003) reported that their self-citation based method outperforms their common citation based method in the experiment (the *Guess-At-Least-One* scenario). As shown in Tables 6-10, however, there are few significant differences between the accuracy with top 10 or fewer guesses by the CC and SC approaches in author identification. Furthermore, the CC approach outperforms the SC approach in affiliation (with top 10 or more guesses) and nationality (with top X or more guesses) identifications. From these results, it is confirmed that not only self-citation but also common citation can be a clue to identify blind entities in a paper. In other words, we need to decrease both of the numbers of self-citations and common citations if we want to increase anonymity of our submitted manuscripts in the blind review process.

Table 11: Cold Start: Identification for top 10 guesses.

Top-10	**Affiliation** [%]				**Nationality** [%]			
	SC	SocC	CC	Full	SC	SocC	CC	Full
MAG(10%)	1.19	0.715	9.42	**9.59**	6.28	3.27	61.8	**62.2**
CS	7.57	2.66	13.9	**15.4**	25.1	5.62	65.8	**66.5**
Eng.	4.18	1.32	9.68	**10.2**	17.1	6.93	62.8	**63.4**
Math	7.03	1.90	16.2	**16.9**	22.8	6.52	76.1	**76.9**
Soc. Sci.	4.78	1.36	6.83	**7.51**	22.8	2.34	59.8	**60.3**

5.2 Identification in Cold Start Scenario

In the previous author identification problem, we can see from Table 3 that approximately 35-70% of test papers in the datasets are not *guessable* as they do not have any link to at least one of the true authors in the training datasets. The affiliations and nationalities in such test papers, however, may be still guessable since other authors who belong to the affiliation and/or other affiliations in the same country may have similar citation history. In this section, we focus on non-*guessable* test papers in the author identification experiment, and guess the true affiliations and nationalities.

In affiliation identification with non-*guessable* papers for author identification, we ignore papers all of whose authors' affiliations are missing in the datasets, and similarly ignore papers in nationality identification all of whose affiliations could not be converted to their counties. As for training, we use the same training datasets and parameters used in Section 5.1. Table 11 shows the performances of our approaches with top 10 guesses and the percentages of guessable papers in affiliation and nationality identifications. The performances of affiliation and nationality identifications in the cold start scenario for author identification are worse than those

Figure 3: Relation between identification rates (top 10 guesses) and author sequence numbers with CS dataset.

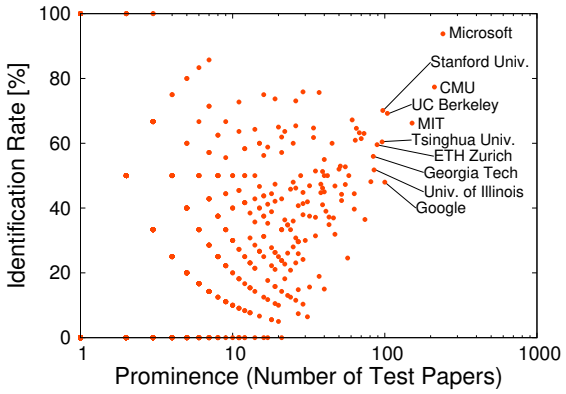

Figure 4: Affiliation prominences and identification rates (top 10 guesses) with CS dataset.

in Tables 6-10. However, at least nationality is still identifiable with a small number of guesses in all the datasets even when we cannot guess true authors in a test paper. Furthermore, we find that the self-citation (SC) model is not useful in this scenario even compared to another baseline approach Rand(S) in nationality identification.

6 Further Analysis

In Section 5, all the entity types are identifiable with a small number of guesses. However, we provide further analysis of the combined model on the CS dataset to answer the following questions.

Which authors are most identifiable?

Figure 3 shows identification rates of different author positions for test papers that have at most 5 authors (85% of the test dataset). As shown, the last author in a paper consistently turns out to be most identifiable, and this may be because the last author is likely to be a director of the research group who may have a stronger research background.

Are prominent affiliations easier to identify?

Here, we consider the number of test papers written by researchers in an affiliation as its prominence. It is apparent from Figure 4 that identification rates of prominent affiliations tend to be high. For example, 93.8% and 77.5% of test papers written by Microsoft and Carnegie Mellon University respectively are identified with top 10 guesses. Note that there are 1,506 affiliations in the graph, but most of the points are overlapped each other.

Are double-blind reviewed papers more anonymized than single-blind reviewed ones?

As shown in Table 12, the performances for papers at single- and double-blind review conferences are almost the same as author and affiliation identifications. This similar performance suggests that the level of anonymity in venues with single-blind review is comparable to that with double-blind review. We only use conferences with at least 40 test papers for denoising here, however, they account for 95% of all test papers.

Table 12: Average percentages of identified papers (top 10 guesses) for single- and double-blind review venues.

CS	**Macro average** [%]		**Micro average** [%]	
Blind review	Single	Double	Single	Double
Author	43.3	42.9	38.3	40.9
Affiliation	55.0	51.9	46.1	48.1

7 Conclusions

The blind review systems are fundamental for research communities to maintain the quality of the published studies. However, it is unclear to what extent the submissions maintain anonymity and how fair the review processes are. In this work, we focus on one of the aspects of de-anonymization by investigating the extent to which we can predict author identity from the paper's citations. Through practical large-scale experiments, we show we can identify author identity, affiliation, and nationality with a few guesses. These results indicate that merely omitting author names is not a sufficient guarantee of anonymity, and may not alleviate fairness considerations in blind review process. This study only involves published papers; analyzing submissions for double-blind review requires considerable involvement of the research communities since they are not public (Tomkins et al., 2017).

Acknowledgements

We thank the anonymous reviewers for their comments. This work is supported in part by a grant from the National Science Foundation (NSF) #IIS-1817183 and #CCRI-1925741. The views in this work do not reflect those of the funding agencies.

References

Judith Gedney Baggs, Marion E. Broome, Molly C. Dougherty, Margaret C. Freda, and Margaret H. Kearney. 2008. Blinding in peer review: The preferences of reviewers for nursing journals. *Journal of Advanced Nursing*, 64(2):131–138.

Homanga Bharadhwaj, Dylan Turpin, Animesh Garg, and Ashton Anderson. 2020. De-anonymization of authors through arxiv submissions during double-blind review.

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

Richard J C Brown. 2007. Double anonymity in peer review within the chemistry periodicals community. *Learned Publishing*, 20(2):131–137.

Cornelia Caragea, Ana Uban, and Liviu P. Dinu. 2019. The myth of double-blind review revisited: ACL vs. EMNLP. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2317–2327. Association for Computational Linguistics.

Wei-Sheng Chin, Yu-Chin Juan, Yong Zhuang, Felix Wu, Hsiao-Yu Tung, Tong Yu, Jui-Pin Wang, Cheng-Xia Chang, Chun-Pai Yang, Wei-Cheng Chang, Kuan-Hao Huang, Tzu-Ming Kuo, Shan-Wei Lin, Young-San Lin, Yu-Chen Lu, Yu-Chuan Su, Cheng-Kuang Wei, Tu-Chun Yin, Chun-Liang Li, Ting-Wei Lin, Cheng-Hao Tsai, Shou-De Lin, Hsuan-Tien Lin, and Chih-Jen Lin. 2013. Effective String Processing and Matching for Author Disambiguation. In *Proceedings of the 2013 KDD Cup 2013 Workshop*, pages 7:1–7:9.

Yuxiao Dong, Hao Ma, Zhihong Shen, and Kuansan Wang. 2017. A Century of Science: Globalization of Scientific Collaborations, Citations, and Innovations. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1437–1446.

Iztok Fister, Iztok Fister, and Matjaž Perc. 2016. Toward the Discovery of Citation Cartels in Citation Networks. *Frontiers in Physics*, 4(December):1–5.

Tao-Yang Fu, Zhen Lei, and Wang Chien Lee. 2015. Patent Citation Recommendation for Examiners. In *Proceedings of the 15th IEEE International Conference on Data Mining*, pages 751–756.

Tao-Yang Fu, Zhen Lei, and Wang-chien Lee. 2016. Modeling Time Lags in Citation Networks. In *Proceedings of the 16th IEEE International Conference on Data Mining*, pages 865–870.

Shawndra Hill and Foster Provost. 2003. The Myth of the Double-Blind Review? Author Identification Using Only Citations. *ACM SIGKDD Explorations Newsletter*, 5(2):179–184.

Amy C Justice, Mildred K Cho, Margaret A Winker, and Jesse A Berlin. 1998. Does Masking Author Identity Improve Peer Review Quality? A randomized controlled trial. *JAMA*, 280(3):240–242.

Dongyeop Kang, Waleed Ammar, Bhavana Dalvi, Madeleine van Zuylen, Sebastian Kohlmeier, Eduard Hovy, and Roy Schwartz. 2018. A dataset of peer reviews (PeerRead): Collection, insights and NLP applications. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1647–1661, New Orleans, Louisiana. Association for Computational Linguistics.

Carole J. Lee, Cassidy R. Sugimoto, Zhang Guo, and Blaise Cronin. 2013. Bias in Peer Review. *Journal of the American Society for Information Science and Technology*, 14(4):90–103.

Michael Levin, Stefan Krawczyk, Steven Bethard, and Dan Jurafsky. 2013. Citation-Based Bootstrapping for Large-Scale Author Disambiguation. *Journal of the American Society for Information Science and Technology*, 14(4):90–103.

Michael J. Mahoney, Alan E. Kazdin, and Martin Kenigsberg. 1978. Getting Published. *Cognitive Therapy and Research*, 2(1):69–70.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Distributed Representations of Words and Phrases and Their Compositionality. In *Proceedings of the 26th International Conference on Neural Information Processing Systems*, volume 2, pages 3111–3119.

Arvind Narayanan, Hristo Paskov, Neil Zhenqiang Gong, John Bethencourt, Emil Stefanov, Eui Chul Richard Shin, and Dawn Song. 2012. On the Feasibility of Internet-Scale Author Identification. In *Proceedings of the 33th IEEE Symposium on Security and Privacy*, pages 300–314.

Mathias Payer, Ling Huang, Neil Zhenqiang Gong, Kevin Borgolte, and Mario Frank. 2015. What You Submit Is Who You Are: A Multimodal Approach for Deanonymizing Scientific Publications. *IEEE Transactions on Information Forensics and Security*, 10(1):200–212.

Alexander M. Petersen, Fengzhong Wang, and H. Eugene Stanley. 2010. Methods for measuring the citations and productivity of scientists across time and discipline. *Physical Review E - Statistical, Nonlinear, and Soft Matter Physics*, 81(3):1–9.

M. Rosen-Zvi, T. Griffiths, M. Steyvers, and P. Smyth. 2004. The Author-Topic Model for Authors and Documents. In *Proceedings of the 20th Conference on Uncertainty in Artificial Intelligence*, pages 487–494.

Arnab Sinha, Zhihong Shen, Yang Song, Hao Ma, Darrin Eide, Bo-June (Paul) Hsu, , and Kuansan Wang. 2015. An Overview of Microsoft Academic Service (MAS) and Applications. In *Proceedings of the 24th International Conference on World Wide Web*, pages 243–246.

Richard Snodgrass. 2006. Single- Versus Double-Blind Reviewing: An Analysis of the Literature. *ACM SIGMOD Record*, 35(3):8–21.

Jessica Su, Ansh Shukla, Sharad Goel, and Arvind Narayanan. 2017. De-anonymizing Web Browsing Data with Social Networks. In *Proceedings of the 26th International Conference on World Wide Web*, pages 1261–1269.

Andrew Tomkins, Min Zhang, and William D. Heavlin. 2017. Reviewer bias in single- versus double-blind peer review. *Proceedings of the National Academy of Sciences*, 114(48):12708–12713.

Alfred Yankauer. 1991. How Blind Is Blind Review? *American Journal of Public Health*, 81(7):843–845.