Trilateral Large-Scale OSN Account Linkability Study

Mishari Almishari
College of Computer and Information Sciences
King Saud University
mialmishari@ksu.edu.sa

Ekin Oguz*
Google, Inc
ekinoguz@google.com

Gene Tsudik
Computer Science Department
University of California, Irvine
gene.tsudik@uci.edu

Abstract

In the last decade, Online Social Networks (OSNs) have taken the world by storm. They range from superficial to professional, from focused to general-purpose, and, from free-form to highly structured. Numerous people have multiple accounts within the same OSN and even more people have an account on more than one OSN. Since all OSNs involve some amount of user input, often in written form, it is natural to consider whether multiple incarnations of the same person in various OSNs can be effectively correlated or linked. One intuitive means of linking accounts is by using stylometric analysis.

This paper reports on one of the first large-scale trilateral stylometric OSN linkability studies.[1] Its outcome has important implications for OSN privacy. The study is trilateral since it involves three OSNs with very different missions: (1) Yelp, known primarily for its user-contributed reviews of various venues, e.g, dining and entertainment, (2) Twitter, popular for its pithy general-purpose micro-blogging style, and (3) Flickr, used exclusively for posting and labeling (describing) photographs. As our somewhat surprising results indicate, stylometric linkability of accounts across these heterogeneous OSNs is both viable and quite effective. The main take-away of this work is that, despite OSN heterogeneity, it is very challenging for one person to maintain privacy across multiple active accounts on different OSNs.

1 Introduction

Online Social Networks (OSNs) have been rapidly gaining worldwide popularity for almost two decades. The OSN paradigm evolved from pre-web BBSs (Bulletin Board Systems) and Usenet discussion groups, through AOL[2] and Yahoo, to enormous and global modern OSNs. One of them, Twitter, has already exceeded $200,000,000$ accounts.[3] In addition to gaining users, OSNs have permeated into many spheres of everyday life. One of many possible ways to classify OSNs is by their primary *mission*:

- **Generic OSNs**, such as Facebook, VK, Google+ and LinkedIn, where users establish and maintain connections while sharing any type of content, of almost any size.
- **Microblogging OSNs**, such as Twitter and Tumblr, that let users share short, frequent and (ostensibly) newsworthy missives.
- **Media-specific OSNs**, such as Instagram and Flickr, where users mainly share content of a certain media type, such as photos or videos. However, even in these OSNs, users provide textual labels and descriptions for shared media content.
- **Review OSNs**, such as Yelp, TripAdvisor and Amazon, where users offer reviews of products and services, e.g. restaurants, hotels, airlines, music, books, etc. These tend to be hybrid sites, that include some social networking functionality, beyond user-provided reviews. Users are evaluated by their reputations and there are typically no size restrictions on reviews.

Despite their indisputable popularity, OSNs prompt some privacy concerns.[4] With growing revenue on targeted ads, many OSNs are motivated to increase and broaden user profiling and, in the process, accumulate large amounts of Personally Identifiable Information (PII). Disclosure of this PII, whether accidental or intentional, can have unpleasant and even disastrous consequences for some OSN users. Many OSNs acknowledge this concern offering adjustable settings for desired privacy levels.

1.1 Motivation

A large number of people have accounts on multiple OSNs, especially, OSNs of different types. For example, it is common for someone in his/her 20-s to have Twitter, Instagram and Facebook accounts. However, privacy **across** OSNs is not yet sufficiently explored. A number of users naturally expect that their accumulated contributions (content) and behavior in one OSN account are confined to that OSN. It would be clearly detrimental to one's privacy if correlating or linking accounts of the same person across OSNs are possible.

*This research was done while the author was at the University of California, Irvine.

[1] One of only two such studies to-date, coincidentally conducted concurrently. The other is (Overdorf and Greenstadt 2016).

[2] http://www.aol.com

[3] https://blog.twitter.com/2013/celebrating-twitter7

[4] This is despite the fact that the entire notion of "OSN Privacy" might seem inherently contradictory.

In this paper, we explore linkability of user accounts across OSNs of different types. That is, given a user holding accounts on two OSNs, we investigate the efficacy and efforts needed to correctly link these accounts. While this problem has been studied in (Goga et al. 2013), prior results are very limited with respect to linkage accuracy and large numbers of accounts. The goal of this work is to develop cross-OSN linkage models that are highly accurate and scalable. To this end, we apply *Stylometry* – the study of one's writing – in a novel framework, that yields very encouraging results. Our linkability study is performed over three popular OSNs: Twitter, Yelp and Flickr. These OSNs are heterogeneous, i.e., each has a very distinct primary mission. Yelp being a community-based review OSN, Twitter is a general purpose microblogging OSN and Flickr is about sharing multimedia content. Thus, the problem of accurately linking users accounts is quite challenging. Figure 1 captures the OSN pairs we study for linkability purposes.

Although accurate and scalable linking techniques are detrimental to user privacy, they can also be useful in forensics, e.g., to trace various miscreants. As is well-known, OSNs have become a favorite global media outlet for both criminals and terrorists to recruit and promote ideology. Both sides of linkability arguments are equally important. However, we believe that it is important to know potential privacy consequences of participating in multiple OSNs, since, as mentioned above, many (perhaps naïvely) expect some confinement or compartmentalization of each OSN account.

1.2 Contributions

Our main anticipated contribution is the Multi-Level Linker Framework (MLLF), a novel idea to hierarchically combine features while scaling the number of possible authors. Using MLLF, we report on the following contributions over literature:

- **High Accuracy.** We develop stylometric-based linkability models that are substantially more accurate than those in previous work with respect to language-based models, e.g, (Goga et al. 2013).
- **Scalability.** Popular OSNs have enormous numbers of users. Thus, scalability of linkability models is essential. Unlike previous work, our models easily scale from 100 to $100,000$ users.
- **Public Data.** Proposed linkability models perform very well with respect to accuracy and scalability even though we assume that the adversary only has access to publicly available textual data from OSNs.[5] Therefore, achieving high accuracy armed only with publicly available data, provides a lower bound on how much the adversary can achieve and serves as an indicator of the severity of the privacy problem.

[5]We believe that users who pursue privacy would disable all OSN meta-data information, such as geo-location – a feature that was essential for linkability accuracy in (Goga et al. 2013). Moreover, private messages will not be available to outside world, which was used in (Afroz et al. 2014).

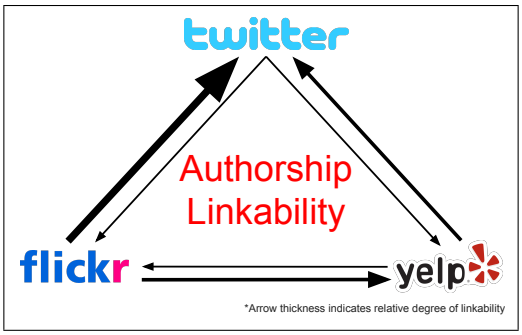


Figure 1: OSN pairs and summary of our linkability study.

2 Related Work

Author Attribution. There has been a lot of research in the field of author attribution. Abbasi, et al. (Abbasi and Chen 2008) proposed a technique based on a new unsupervised learning method, referred to as Writeprints. It uses Karhunen-Loeve transforms along with a rich set of features to identify authors, achieving accuracy of 91% in finding the author of an anonymous message from a set of 100 candidate authors. A study called Herbert West – Deanonymizer, was conducted to investigate the possibility of de-anonymizing peer reviews of academic papers (Nanavati et al. 2011). A high percentage – around 90% – of reviews were correctly de-anonymized from a set of 23 reviewers using Naïve Bayes Classifier. Another recent effort studied author identification of the Internet blogs on a relatively large-scale, with $100,000$ authors (Narayanan et al. 2012). In certain cases, de-anonymization accuracy of 80% was achieved and anonymous texts were linked across different platforms. Mishari, et al. (Mishari and Tsudik 2012) studied linkability of community-based reviews in Yelp, based on a set of about $2,000$ reviewers and almost all reviews were correctly de-anonymized. Even though a simple feature set was used (e.g., unigrams and bigrams) with Naïve Bayesian classifier, high linkability accuracy was achieved. Stamatatos (Stamatatos 2009) extensively surveyed the area of author attribution and we refer to it for a good overview of the topic.

Cross-Linking Accounts. The study most relevant to this paper was conducted by Goga, et al. (Goga et al. 2013). It cross-linked accounts between different OSNs, the same three that are used in this paper: Twitter, Yelp and Flicker. Features that included locations, timestamps and text were used, with the help of the cosine distance function, to link accounts operated by the same user across OSNs. While the setting is similar to ours[6], we substantially improve linkability results with respect to language-based models. Unlike (Goga et al. 2013), we rely only on text-based features and leverage them to improve scalability (larger set of accounts) and linkability results. Moreover, we report on correlations in between all OSN pairs, whereas (Goga et al. 2013) only discusses correlating Yelp and Flickr to Twitter. Further comparisons on the performance are reported in

[6]As acknowledged in Section 5, we borrowed our dataset from this study.

Section 7. A follow-up work by Goga, et al. (Goga et al. 2015) explored the reliability of profile matching by showing the performance degradation of matching when a more reliable sample set – similar to original OSN's with respect to the large number of false matches– was used. Additionally, several optimizations were proposed that minimally improved matching. While our matching tests can be extended by adding more false matches (possible future work), we still believe that our technique is heavily tested under different set sizes (up to $100,000$). Having sampled sets similar to the original OSN's is not always necessary to emulate real-world privacy attacks as attackers may rely on some other techniques (possibly manual) to narrow down the list and exclude false matching accounts

Similarly, Afroz, et al. (Afroz et al. 2014) successfully explored cross-linking multiple accounts belonging to the same user within the same forum or blog-based site. This is a step forward since, in prior studies, linking was based on artificially created accounts of the same user. Accuracy between 85% and 90% was achieved, while maintaining high recall values. The study used an algorithm called **Doppelgänger Finder**, where two accounts: $account_A$ and $account_B$ were claimed to belong to the same user if combined probability of attributing $account_A$ to $account_B$ and vice versa exceeded a specific threshold. The probability of attributing $account_A$ to $account_B$ was computed based on a model trained on all accounts except $account_A$ and vice versa. Probabilities are combined by averaging, multiplying or square-averaging. Lexical, domain and syntactic features were used along with Principal Component Analysis to reduce the feature set size.

A large-scale $(10,000)$ author attribution study was recently conducted to link Twitter accounts based on very simple lexical features – unigrams and bigrams – and Naïve Bayesian classifier (Almishari et al. 2014). High linkability results – nearly 100% – have been achieved. Also, results were verified based on *ground truth* – actual Twitter accounts that belong to the same user.

Other related work explored account linkability in online services based on entropy of user-names (Perito et al. 2011). In (Irani et al. 2009), account properties with a simple set of heuristics were used to cross-link users. And finally, Iofciu, et al. explored tags to identify users across Delicious, StumbleUpon and Flickr (Iofciu et al. 2011).

Concurrently with the research presented in this paper, cross-domain stylometric experiments between Blogs, Twitter feeds and Reddit comment were conducted in (Overdorf and Greenstadt 2016). With about $10,000$ words per author to train the model, this work attains accuracy of up to 70%. However, evaluation in (Overdorf and Greenstadt 2016) comprises only 50 authors, since dealing with larger numbers of authors turns out to be computationally very expensive.

3 OSN Background

In this section, we overview three OSNs used in our study.

Yelp is a community-based review site where users – who must have accounts – offer reviews of various products and services. Access to reviews is not restricted, i.e., anyone can read Yelp reviews, with or without an account. Typical reviewed industry categories include: restaurants, automotive, medical, hospitality and entertainment. At least in North America, Yelp is very popular: the number of reviews exceeds $70,000,000$ and the number of yearly visitors is about $142,000,000$ (yel a). Yelp is considered to be an OSN since it also allows its users to connect to, and interact with, other Yelp users. Yelp has a reward system for reviewers based on the quantity and quality (popularity and ratings) of their contributions. Not surprisingly, this helps increase the number of avid or prolific reviewers (yel b).

Twitter is a microblogging OSN where registered users (known as tweeters) post short messages (called tweets).[7] Some tweeters make their tweets public, meaning that anyone can read them regardless of having a Twitter account. Meanwhile, others restrict access to their tweets to so-called followers – Twitter users who have explicitly requested, and have been granted, access to one's tweets. One of Twitter's most distinctive features is the 140-character size limit for tweets. Twitter is currently one of the most popular and diverse OSNs, having attracted many avid tweeters among politicians, journalists, athletes and various celebrities. Furthermore, all kinds of groups, societies and organizations (both in public and private sectors) have strong Twitter presence. The number of Twitter accounts exceeds $200,000,000$ (twi).

Flickr is a focused OSN and a cloud storage provider, specializing in sharing multimedia content, i.e., photographs and videos. Flickr users can annotate their multimedia content with text. Without annotations, the file-name of a particular photo or video content is used as a default title. Unlike Twitter, Flicker imposes no size limit on the annotation text. Using Flickr to post (or view restricted) content, generally requires having an account. However, public content can be viewed by anyone. Flickr has a notion of a *contact*, akin to a friend or a connection on other OSNs.

As follows from the above description, each of these three OSNs is quite distinct in its primary mission. This makes the problem of linking accounts across them particularly challenging.

4 Problem Setting

The author attribution problem can be informally defined as:

> *Given a set of known authors* $A_{known} = \{a_1, a_2, ..., a_n\}$*, and an anonymous contribution* C *(textual, non-textual or a mix of both), find the most likely candidate author of* C *among those in* A_{known}*.*

In the OSN context, author attribution problem translates into finding the most likely candidate author of anonymous posts, i.e., the user who most likely generated these posts given his or her OSN profile. We refer to attribution of anonymous posts to a user account as *linking*.

As mentioned earlier, our goal is to study the author attribution problem (based on stylometry) across multiple OSNs.

[7] Technically, one can be a Twitter user but not a tweeter, e.g., someone might create an account only to follow others' tweets, but not tweet.

Basically, we assume that some people have accounts in two OSNs and we want to link these accounts. We have OSN_1 and OSN_2 each with its own set of accounts. We first remove from each OSN accounts that do not have a match (authored by the same user) in the other OSN. This results in R_OSN_1 and R_OSN_2 that are reduced versions of OSN_1 and OSN_2, respectively. To make the problem more challenging and also more realistic, we pollute R_OSN_2 by introducing additional X randomly chosen accounts that were originally in OSN_2. As a result, for each account in R_OSN_1, there is a matching account in R_OSN_2. We refer to the accounts in R_OSN_1 as *unknown*, and those in R_OSN_2 – as *known*, accounts.

Now, the problem is reduced to finding a matching model M, i.e., an author attribution technique, that links unknown accounts in R_OSN_1 to known accounts in R_OSN_2. Specifically, for each unknown account in R_OSN_1, M returns a list of all accounts in R_OSN_2 sorted in decreasing order of probability of the correct match. Similar to prior work in (Mishari and Tsudik 2012), we define Top-K linkability ratio –LR– of M as the ratio of unknown accounts (accounts in R_OSN_1) that have their correct matching account – in R_OSN_2 – among the Top-K accounts of their returned lists from M. Our goal boils down to finding a matching model that maximizes LR with respect to X and K. We vary X so that the total number of known accounts ranges from 100 to $100,000$. Furthermore, we vary K among 1, 10 and 100 [8].

5 Dataset

We use the base dataset obtained (crawled) and used by Goga et al. (Goga et al. 2013). Encompassing users from Yelp, Twitter and Flickr, this dataset is gigantic, containing over $350,000,000$ tweets, $29,000,000$ Flickr posts and $1,000,000$ Yelp reviews. Its most important property is the ground truth of matching accounts: it provides a set of users who operate accounts in multiple OSNs. In the rest of this section, we describe the data cleaning process and then provide more details regarding matching accounts.

5.1 Data Cleaning

Our initial analysis of the base dataset revealed the existence of numerous users with very limited overall contributions. However, stylometric analysis is known to perform accurately in the context of highly prolific users. Some recent studies, (Afroz et al. 2014), (McDonald et al. 2012) and (Rao and Rohatgi 2000), report achieving good linkability performance with at least $4,500$ words per author. Thus, we first need to cull users with lower overall contributed text. We also need to filter out contributions that did not originate with the target user, since some OSNs (e.g., Twitter) allow users to repost (re-tweet) what other people have posted. This filtering helps us better capture users' own stylometric properties. Consequently, we filter out Twitter re-tweets, URLS, user mentions, and posts in languages other than English.

After filtering, we combine all remaining posts of users into a single body of text. This corresponds to the union of Yelp reviews, Twitter tweets and Flicker photo annotations. As the last step, we remove all users who have a cumulative word count of less than $1,000$ and we normalize profile vector according to word count. We stress that this threshold of only $1,000$ words per author is significantly lower than that in previous studies, e.g., $4,500$ words in (Afroz et al. 2014), (McDonald et al. 2012) and (Rao and Rohatgi 2000).

5.2 Matching Accounts

Dataset includes a set of matching accounts that correspond to what we refer to as: *ground truth*. This set links usernames from different OSNs. This information was collected in (Goga et al. 2013) using the "Friend Finder" functionality provided in OSNs. Friend Finder was run on input of a list of $10,000,000$ e-mail addresses using browser automation tools: Watir and Selenuim[9]. Then, the list of users registered with the given e-mail addresses was checked, in order to identify user-names registered to the same e-mail address, i.e. operated by the same person.

After data cleaning, a sufficient number of matching accounts remain for linkability experiments: 153 for Yelp-Twitter, 299 for Twitter-Flickr and 55 Yelp-Flickr.

6 Preliminaries

Before presenting experimental results, we provide some background information about the feature set and the methodology.

6.1 Feature Set

We construct a unique set of features, using a subset of the popular Writeprints set (Abbasi and Chen 2008) along with 3 additional features. Writeprints contains 22 distinct stylometric features. From these, we select 9 lexical, syntactic and content features before adding 3 more custom features (not present in Writeprints). The resulting 12 features are:

- **Lexical** features include frequencies of alphabetical n-grams (n consecutive letters) and special characters, e.g. "*", "@", "#", "$", and "%".
- **Syntactic** features consist of frequencies of function words, punctuation characters and Part-Of-Speech (POS) tags, where unigrams correspond to one tag, and bigrams to two consecutive tags. Function words are 512 common function words used by Koppel et al. (Koppel, Schler, and Zigdon 2005).
  POS tags are grammatical descriptions of words in sentences, e.g, adjective, noun, verb and adverb. We use two popular POS taggers:
  1. Stanford Log-linear (Toutanova et al. 2003), which was the booster of account linkability in recent studies (Afroz et al. 2014; Almishari, Oguz, and Tsudik 2014).

[8] Note that in our model, we can still compute the True Positive and False Positive rates by marking the top matches (the top of the returned lists) as True and others as False.

[9] Watir: `https://github.com/watir/watir` and Selenium: `http://www.seleniumhq.org/`

2. GATE Twitter (Derczynski et al. 2013), which has never been used in account linkability before.

POS tagging of tweets is hard due to the short message style in Twitter. Therefore, we integrate GATE – a state-of-art accurate POS tagger specially designed for Twitter – to our feature set. Our experimental results demonstrate that GATE Twitter tagger improves the account linkability significantly.

- **Content** features include frequency of words. This is the only stylometric feature used in (Goga et al. 2013) for linking accounts.

Features are computed for each user profile. Each feature is normalized by the total count of features within the same category.

Similar subsets of Writeprints were used in several prior linkability studies, e.g., Afroz et al. (Afroz et al. 2014; Almishari, Oguz, and Tsudik 2014), to yield high linkability accuracy. Encouraging results using Letter Quads (4-grams) are achieved in Kevselj et al. (Kešelj et al. 2003). To the best of our knowledge, GATE Twitter POS features have never been used in linkability studies before.

6.2 Methodology

Based on the setting described in Section 4, we have two sets of accounts: known and unknown. We want to accurately match unknown accounts to their known counterparts, while maintaining the highest possible Top-K Linkability Ratio (LR). For that, we first convert each user profile into a feature vector: $F_T = \{F_{T_1}, F_{T_2}, ..., F_{T_n}\}$ where F_{T_i} denotes the i-th token for feature F_T.

Next, we initiate a distance learning model using Chi-Square Distance (CS_d) to link an unknown account to a known one. Specifically, for each unknown account a_u, we calculate the $CS_d(a_u, a_{k_j})$ where j varies over all possible known accounts. Finally, we rank the distances in ascending order and output the resulting ordered list, where the first entry represents the most likely match of the known account a_k to the unknown account a_u.

7 Experimental Results

This section presents the results of the large-scale trilateral OSN account linkability study. We begin with the baseline result. Next, we outline the new Multi-Level Linker Framework which significantly improves on the baseline. Then, we show how this framework yields scalable linkability ratios (LRs) for up to $100,000$ authors. Finally, we present and discuss experiment execution times & memory footprint.

7.1 Baseline

Using the methodology from the previous section, we experiment with various features. Similar to prior work in (Almishari, Oguz, and Tsudik 2014), we apply a greedy hill-climbing algorithm to assess the effects of every feature. We start with all features individually. Then, we combine the best-performing features and assess the amount of improvement. We present the baseline assessment only for Yelp↔Twitter linkability, since other sets perform similarly. Following Section 4, we set the list of unknown accounts $A_{unknown}$ to the full-set of matching accounts as (153 accounts) while we set the size of the known accounts A_{known} to $1,000$.

Table 1 shows Top-1 LRs of individual features. At best, Yelp→Twitter already shows a relatively high 55% Top-1 LR, while Twitter→Yelp performs quite poorly, at 10%.

Feature Index	Twitter→Yelp	Yelp→Twitter
1: Letter Uni	1%	1%
2: Letter Bi	1%	43%
3: Letter Tri	**7%**	**55%**
4: Letter Quad	**10%**	**53%**
5: Special Chars	1%	0%
6: Func. Words	3%	**50%**
7: Punc. Marks	0%	1%
8: Stanford POS Tags Uni	1%	8%
9: Stanford POS Tags Bi	3%	27%
10: Words	**9%**	39%
11: GATE POS Tags Uni	2%	7%
12: GATE POS Tags Bi	3%	18%

Table 1: Top-1 LRs using the baseline Chi-Square methodology. Boldfaced cells represent the highest LRs.

Next, we combine the best features (highlighted in boldface) from Table 1 and show improved results in Table 2.

Features	4&10	3&10	3&4	3&4&10
Twitter→Yelp	**11%**	**8%**	**9%**	**9%**

Features	3&4	3&6	4&6	3&4&6
Yelp→Twitter	**54%**	**59%**	**57%**	**56%**

Table 2: Top-1 LRs, with combined best features from Table 1.

For the Twitter→Yelp case, when Letter Quadgrams and Words features are combined, results are slightly better than the baseline. However, after combining more than two features, we observe a decrease in LR. As for Yelp→Twitter, LR increases slightly when best features are combined (3&6). Similar to Twitter→Yelp, combining more than two features decreases LR.

These results are comparable to those obtained in language-style correlation investigated in (Goga et al. 2013). Likewise, we achieve modest LRs, even with more complex language-based features. To summarize, recent techniques that work reasonably well within the same OSNs, do not appear to be as effective across OSNs. To this end, in the next section, we construct the Multi-Level Linker Framework, which, according to our experiments, significantly boosts linkability.

7.2 Multi-Level Linker Framework (MLLF)

While experimenting with various combination of features, we notice that combining many of them increases noise and prolongs run-times especially when number of authors increases. Moreover, dimensionality reduction techniques like SVD, do not help increase linkability. Similar observations

about experimental run-time infeasibility are also mentioned in (Afroz et al. 2014) and (Goga et al. 2013). This motivates us to explore how to make better use of all textual features, in order to scale for many authors.

We now present the Multi-Level Linker Framework (MLLF), a novel idea to combine features in a scalable manner. To our knowledge, an algorithm similar to MLLF is not available in the literature.

The intuition behind MLLF is the use of features in a more hierarchical manner while all features contribute to the result. The basic idea is to run linkability experiments at multiple levels, with each level using a different feature category. After each level, we halve the number of known authors, for every unknown author. This is done by filtering out the most distant (least likely) known authors. Then, at the next level, we use a different feature category with the remaining known authors. We apply this technique for every feature category, and eventually —-after progressing through all the features—- output the final top position of the matching account. In every experiment, we randomly permute the order of feature categories. We run experiments in 10-fold and report the averages of final linkability results. In plots, we provide positive and negative error bars to average linkability results in order to better understand the effects of feature ordering. Since there is no clear way of feature ordering, (and trying all the permutations to select the best one will not scale for many authors), we pick a random order and leave optimizing the feature ordering to future work. High-level pseudocode for MLLF can be found in the Appendix of (Almishari, Oguz, and Tsudik 2015).

Applying MLLF yields significantly higher LRs with respect to the baseline. Improvements – between [27%, 73%] – in Top-1 LR, when the number of known authors is $1,000$, are:

Twitter→Yelp	11% → 63%
Yelp→Twitter	59% → 88%
Twitter→Flickr	11% → 54%
Flickr→Twitter	67% → 94%
Flickr→Yelp	13% → 86%
Yelp→Flickr	5% → 66%

7.3 Scalability: Number of Authors

Having obtained an improvement over baseline results, we now consider MLLF's scalability. To this end, we vary the number of known authors from 100 to $100,000$ and examine how LRs are affected.

From 100 **to** $1,000$ In the first batch, we experiment with $|A_{known}|$ from 100 to $1,000$. OSN pairs with the highest Top-1 LRs are shown in Figure 2a. OSN→Twitter LRs gets as high as 95% while OSN→Yelp LRs gets 90% in a set of $1,000$ authors. We notice linkability to Twitter is higher than linkability to Yelp in all cases. Also, when number of author increases, OSN→Yelp LRs decreases more than OSN→Twitter. Lastly, OSN→Yelp linkability results shows higher variance, that is affected more by the order features.

	Top-1		*Top-10*	
Number of Authors	*100*	*1,000*	*100*	*1,000*
Yelp→Flickr	77%	73%	93%	92%
Twitter→Flickr	65%	63%	88%	89%

Table 3: Top-1 and Top-10 LRs of OSN→Flickr as the number of authors grows from 100 to $1,000$

OSN→Flickr exhibits the worst results; LRs are shown in Table 3. Top-1 LR of Twitter→Flickr drops to 63% in a set of $1,000$ authors. Interestingly, LRs of OSN→Flickr does not decrease as much as OSN→Yelp. While Top-1 LRs of OSN→Yelp decreases as much as 15%, the biggest decrease is only 4% for OSN→Flickr when number of authors grows from 100 to $1,000$.

From $1,000$ **to** $10,000$ Next, we vary the number of authors from $1,000$ to $10,000$. (The actual number of accounts in $Yelp'$ is $9,348$, which we round to $10,000$ to simplify the graphs.)

Firstly, Top-10 LRs of OSN→Yelp and and OSN→Flickr are shown Figure 2b. We observe that Flickr→Yelp LR achieves 89% and Twitter→Yelp achieves 80% in Top-10. In contrast, Yelp→Flickr is 72% and Twitter→Flickr is 70%. Also, OSN→Yelp is more resilient to random feature ordering than OSN→Flickr. Furthermore, both Yelp and Twitter perform very similarly when linking to a Flickr account.

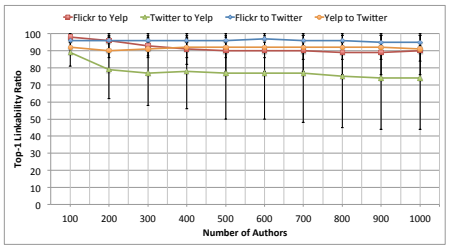
Secondly, Table 4 summarizes linkability results for all OSN combinations. Top-1 LR for $10,000$ authors drops to as low as 29% in Yelp→Flickr, and grows as high as 86% in Flickr→Twitter. Similar to trends in Section 7.3, the highest Top-1 LRs among all OSN combinations is 86% for Flickr→Twitter, followed by 77% for Yelp→Twitter when the number of authors is $10,000$. Moreover, OSN→Twitter model continues to show low linkability variance – 6% in Flickr→Twitter and 9% in Yelp→Twitter – according to the order of features.

For Top-1, linkability to Twitter is best, while linkability to Flickr is worst. For Top-10, the results are really encouraging with 70% as the lowest LR for a set of $10,000$ authors. Lastly, linkability to Twitter decreases by only 2% when number of authors changes from $1,000$ to $10,000$.

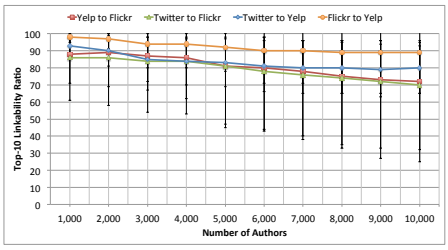
	Top-1		*Top-10*	
Number of Authors	*1,000*	*10,000*	*1,000*	*10,000*
Flickr→Twitter	94%	86%	98%	97%
Yelp→Twitter	88%	77%	99%	97%
Flickr→Yelp	86%	63%	98%	89%
Twitter→Yelp	63%	45%	93%	80%
Yelp→Flickr	66%	29%	88%	72%
Twitter→Flickr	54%	38%	86%	70%

Table 4: Top-1 and Top-10 LRs when # of authors grows from $1,000$ to $10,000$

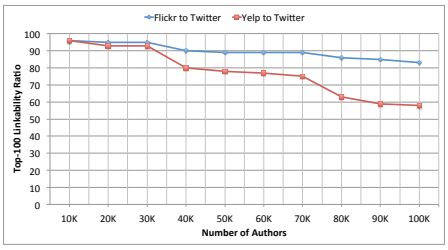
From $10,000$ **to** $100,000$ As the final step in the scalability exercise, we increase $|A_{known}|$ to $100,000$ authors. Only $Twitter$ has up to $100,000$ authors after cleaning. Thus,



a: Top-1 LRs of OSN→Twitter and OSN→Yelp; 100 to 1,000



b: Top-10 LRs of OSN→Yelp and OSN→Flickr; 1,000 to 10,000



c: Top-100 LRs of OSN→Twitter; 10,000 to 100,000

Figure 2: LRs when number of authors grows from 1,000 to 100,000

we only experiment with Flickr→Twitter and Yelp→Twitter combinations. Also, we remove Letter Quadgrams from the feature set and run this batch of experiments with the remaining 11 features, due to memory problems experienced with over 90,000 authors.

Figure 2c shows Top-100 LRs and Table 5 shows Top-1 and Top-10 LRs. Notably, even in the extreme case of 100,000 authors, we can still link to the known author with 54% accuracy in Flickr→Twitter, and 18% accuracy in Yelp→Twitter. If we relax the linkability goal to Top-100, Flickr→Twitter grows to 83% and Yelp→Twitter to 58%. We notice that linkability from Flickr is higher than that from Yelp. Moreover, the former is less affected by the increase in the number of authors: Flickr→Twitter Top-1 LR decreases by 26% while Yelp→Twitter decreases by 50%.

	Top-1		*Top-10*	
Number of Authors	*10,000*	*100,000*	*10,000*	*100,000*
Flickr→Twitter	80%	54%	91%	68%
Yelp→Twitter	68%	18%	88%	42%

Table 5: Top-1 and Top-10 LRs as # of authors grows from 10,000 to 100,000.

7.4 Execution Time and Memory Footprint

Scalability in real-world OSNs begins with at least several millions of users. Therefore, it is very important to assess performance of a linkability study (such as ours) in order to test whether it is applicable in the real world.

We ran all experiments on a 64-processor machine: Intel(R) Xeon(R) CPU E5-4610 v2 @ 2.30GHz, with 128GB of memory. Multi-threaded experiment code is implemented in Java and executed under Ubuntu 14.04 LTS. We used MongoDB to store and query the datasets. Note that all the features are precomputed and saved to this database. This saves us a tremendous amount of execution time, since feature extraction becomes very time-, memory- and storage-consuming, especially, for dynamic features such as Words and Part-of-Speech Tags. We plan to make all of the source code publicly available prior to publication of this paper.

Run-time complexity of the MLLF algorithm (to link a single unknown account) is $O(|A_{known}| * CS_d * |F|)$. This complexity is proportional to size of the known accounts set,

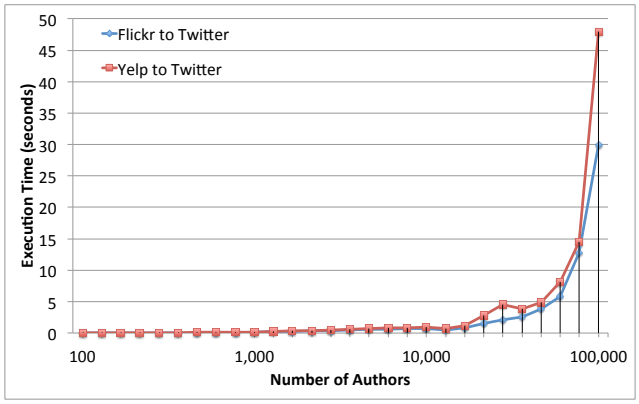


Figure 3: Execution times (seconds) of MLLF with variable # of authors from 100 to 100,000

time to calculate Chi-Square distance between two feature sets and number of feature categories.

Figure 3 shows execution times when $|A_{known}|$ varies from 100 to 100,000. We observe linear trend, as expected from the algorithm complexity. Execution time reaches almost 1 second for 10,000 authors, and approximately 13 seconds for 90,000 authors. We observe an exponential jump for 100,000 authors This occurs because of insufficient RAM, which forces the code to resort to using the disk swap partition.

After the execution times, we present the memory footprint of MLLF in Figure 4. Since running MLLF with more than 90,000 authors causes disk swap partition usage, we are only showing memory consumption up to 80,000 authors. As expectedly, memory usage increases linearly while author set size grows. MLLF requires 7 gigabyte of memory for 1,000 authors, 24 gigabyte for 10,000 authors and 111 gigabyte for 80,000 authors. Most important memory characteristics of MLLF is even though algorithms work in hierarchical increments, memory usage does not increase after each level. This is because MLLF is using only one feature category in each level. Thus, conventional algorithms, that uses more than one feature category, would require a lot more memory than MLLF.

Of course, better software engineering practices would likely lower the memory footprint and improve execution

time. However, we believe that current results give a general idea of MLLF's scalability. For example, in only 13 seconds, MLLF can link an unknown account with 71% accuracy, within a set of $90,000$ authors.

7.5 Summary

Our experimental results can be summarized as follows:

1. We begin with a baseline method using a greedy hill-climbing algorithm on features to improve linkability. This results in 11% Top-1 LR from Twitter→Yelp, which is comparable to prior results in (Goga et al. 2013). We concluded that recent stylometric linkability models are not resilient when used to link accounts across heterogeneous OSNs; see Section 7.1.
2. We then proposed a new Multi-Level Linker Framework (MLLF), which improves LRs by around 50%; see Section 7.2.
3. Next, we demonstrated MLLF's scalability when the number of authors grows from 100 to $100,000$. We managed to reach Top-10 LRs of 68% for Flickr→Twitter and 42% for Yelp→Twitter in a set of $100,000$ possible authors; see Section 7.3.
4. Finally, we discussed the run-times and memory requirements of MLLF as the number of authors increases. MLLF only takes around 8 seconds to link an unknown account from either Flickr or Yelp to Twitter in a set of $80,000$ possible authors, and requires around 111 gigabyte of memory; see Section 7.4.

Linkability Improvement over Goga, et al: Our results significantly improve on the prior work of Goga, et al (Goga et al. 2013) with respect to language-based models. Even though their setting is slightly different from ours (we perform data cleaning and filtering of low-prolific users), we achieve True Positive Rate of 60% in Flickr→Twitter and 36% in Yelp→Twitter in a set of 70,000 authors (with negligible false positive ratios), while (Goga et al. 2013) reaches 13% for the former and 9% for the latter using language profile in a set of $75,747$ authors.[10] When using other features (username and location), models in (Goga et al. 2013) outperform ours. However, we believe that matching techniques based on such features are easily defeatable – e.g. locations can be disabled and usernames are changeable. Moreover, we experiment and report linkability ratios from all OSN pairs while (Goga et al. 2013) only experiments with OSN→Twitter. In the next section, we discuss the results in more detail.

8 Discussion & Future Work

We now attempt to elaborate on some potential issues and future work prompted by the results described in the paper.

Our initial and somewhat intuitive expectation was that linkability to Yelp would be the highest, since Yelp, unlike Twitter, does not have text size limits. We anticipated that a typical Yelp user exhibits a writing style very similar to that used in their everyday writing activities. In contrast, Twitter forces certain verbal contortions and compressions due to its 140-characters limitation. However, it turns out that Twitter

[10]We set Top-1 LRs as True Positive Rate.

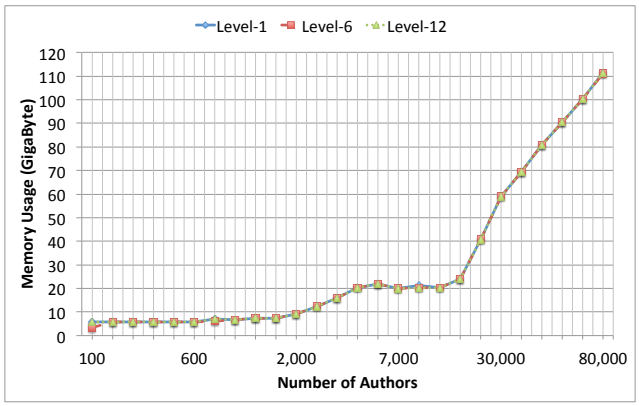


Figure 4: Memory footprint of MLLF running for Flickr→Twitter (memory consumptions is similar in other OSN combinations) when number of authors increases from 100 to $80,000$. Each curve refers to a different level in MLLF.

allows us to build a better stylometric profile than Yelp. One potential explanation is restricted context or focus: Twitter is a general-purpose micro-blogging OSN, while Yelp is primarily about reviewing restaurants, hotels and various other venues. In Twitter, people write mostly about themselves, other people, events (e.g., news), yet the context is totally unrestricted, i.e., anything goes. This could mean that contextual freedom allows capturing one's writing style better as long as a user authors a sufficient overall amount of text.

MLLF's complexity increases linearly with the number of accounts. Therefore, we believe it can be used in a much bigger account set, given enough RAM. According to the trend observed in our experiment execution times, we estimate that it would take around 2.5 minutes to link one unknown account to $1,000,000$ known ones. Of course memory footprint, multi-threading and implementation efficiency can be further optimized using better software engineering practices, which we also leave to future work. Moreover, current implementation of MLLF shuffles available features and uses a different feature in each level. One can imagine that if a feature is weak and is unfortunately chosen in early levels, then the true match will be filtered out. As part of our future work, we plan to investigate how to order features so that linkability will be maximized. Additionally, Extending MLLF's feature set with other Writeprints features is very likely to influence LRs. As part of future work we plan to gradually experiment with other Writeprints features.

We do not yet know how combining homogeneous and/or heterogeneous accounts influences linkability. This is another open question. One obvious step is to combine Yelp and Twitter profiles of known accounts, while trying to link to an unknown Flickr account. Such a hypothetical system could generate a generic stylometric fingerprint, which would be a real breakthrough in author attribution and linkability. We leave exploring this to future work.

9 Conclusions

Despite the elusiveness of OSN privacy, many users expect that multiple accounts they operate within one, and on more than one, OSNs remain isolated, i.e., unlinkable, owing perhaps to very different OSN missions. For example, photo-sharing, micro-blogging and product/service reviews appear to be quite distinct types of OSN specialization. However, this is unfortunately not the case, as supported by the results of the study presented in this paper. It also represents the first large-scale stylometric-based account linkability experiment conducted across three heterogeneous OSNs: Yelp, Twitter and Flickr.

ACKNOWLEDGMENTS

We are very grateful to the authors of (Goga et al. 2013) for kindly sharing the crawled Yelp, Flickr, and Twitter datasets used in their previous work. This research was supported in part by the NSF Award 1212943: "CSR: Collaborative Research: Enabling Privacy-Utility Trade-Offs in Pervasive Computing Systems."

References

Abbasi, A., and Chen, H. 2008. Writeprints: A stylometric approach to identity-level identification and similarity detection in cyberspace. *ACM Transactions on Information Systems (TOIS)* 26(2):7.

Afroz, S.; Islam, A. C.; Stolerman, A.; Greenstadt, R.; and McCoy, D. 2014. Doppelgänger finder: Taking stylometry to the underground. In *Security and Privacy (SP), 2014 IEEE Symposium on*, 212–226. IEEE.

Almishari, M.; Kaafar, D.; Oguz, E.; and Tsudik, G. 2014. Stylometric Linkability of Tweets. In *WPES*.

Almishari, M.; Oguz, E.; and Tsudik, G. 2014. Fighting authorship linkability with crowdsourcing. In *Proceedings of the second edition of the ACM conference on Online social networks*, 69–82. ACM.

Almishari, M.; Oguz, E.; and Tsudik, G. 2015. Trilateral large-scale osn account linkability study. *arXiv preprint arXiv:1510.00783*.

Derczynski, L.; Ritter, A.; Clark, S.; and Bontcheva, K. 2013. Twitter part-of-speech tagging for all: Overcoming sparse and noisy data. In *RANLP*, 198–206.

Goga, O.; Lei, H.; Parthasarathi, S. H. K.; Friedland, G.; Sommer, R.; and Teixeira, R. 2013. Exploiting innocuous activity for correlating users across sites. In *Proceedings of the 22nd international conference on World Wide Web*, 447–458. International World Wide Web Conferences Steering Committee.

Goga, O.; Loiseau, P.; Sommer, R.; Teixeira, R.; and Gummadi, K. 2015. On the reliability of profile matching across large online social networks. In *KDD*.

Iofciu, T.; Fankhauser, P.; Abel, F.; and Bischoff, K. 2011. Identifying users across social tagging systems. In *ICWSM*.

Irani, D.; Webb, S.; Li, K.; and Pu, C. 2009. Large online social footprints–an emerging threat. In *CSE '09: Proceedings of the 2009 International Conference on Computational Science and Engineering*, 271–276. Washington, DC, USA: IEEE Computer Society.

Kešelj, V.; Peng, F.; Cercone, N.; and Thomas, C. 2003. N-gram-based author profiles for authorship attribution. In *Proceedings of the conference pacific association for computational linguistics, PACLING*, volume 3, 255–264.

Koppel, M.; Schler, J.; and Zigdon, K. 2005. Automatically determining an anonymous author's native language. In *Intelligence and Security Informatics*. Springer. 209–217.

McDonald, A. W.; Afroz, S.; Caliskan, A.; Stolerman, A.; and Greenstadt, R. 2012. Use fewer instances of the letter "i": Toward writing style anonymization. In *Privacy Enhancing Technologies*, 299–318. Springer.

Mishari, M. A., and Tsudik, G. 2012. Exploring linkability of user reviews. In *ESORICS*.

Nanavati, M.; Taylor, N.; Aiello, W.; and Warfield, A. 2011. Herbert West – Deanonymizer. In *6th USENIX Workshop on Hot Topics in Security*.

Narayanan, A.; Paskov, H.; Gong, N. Z.; Bethencourt, J.; Stefanov, E.; Shin, E. C. R.; and Song, D. 2012. On the Feasibility of Internet-Scale Author Identification. In *IEEE Symposium on Security and Privacy*.

Overdorf, R., and Greenstadt, R. 2016. Blogs, twitter feeds, and reddit comments: Cross-domain authorship attribution. *Proceedings on Privacy Enhancing Technologies* 2016(3):155–171.

Perito, D.; Castelluccia, C.; Kaafar, M. A.; and Manils, P. 2011. How Unique and Traceable Are Usernames? In *PETS*.

Rao, J. R., and Rohatgi, P. 2000. Can pseudonymity really guarantee privacy? In *USENIX Security Symposium*.

Stamatatos, E. 2009. A Survey of Modern Authorship Attribution Methods. In *Journal of the American Society for Information Science and Technology*.

Toutanova, K.; Klein, D.; Manning, C. D.; and Singer, Y. 2003. Feature-rich part-of-speech tagging with a cyclic dependency network. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology-Volume 1*, 173–180. Association for Computational Linguistics.

Twitter Blog. https://blog.twitter.com/2013/celebrating-twitter7. Last accessed on 2015-04-26.

Yelp – About Us. http://www.yelp.com/about.

Yelp Elite Squad. http://www.yelp.com/elite. Last accessed on 2015-04-23.