Stylometric Linkability of Tweets

Mishari Almishari
King Saud University
mialmishari@ksu.edu.sa

Dali Kaafar
NICTA
dali.kaafar@nicta.com.au

Gene Tsudik Ekin Oguz
UC Irvine
{gtsudik,eoguz}@uci.edu

ABSTRACT

Microblogging is a very popular Internet activity that informs and entertains a large number of people world-wide via quickly and scalably disseminated terse messages containing all kinds of news-worthy utterances. Even though microblogging is neither designed nor meant to emphasize privacy, numerous contributors hide behind pseudonyms and compartmentalize their different incarnations via multiple accounts within the same, or across multiple, site(s).

Prior work has shown that stylometric analysis is a very powerful tool capable of linking product or service reviews and blogs that are produced by the same author, even when the number of authors is large. In this paper, we explore linkability of tweets. Our results, based on a very large corpus of tweets, clearly demonstrate that, at least for relatively active tweeters, linkability of tweets by the same author is easily attained even with a large number of tweeters. Furthermore, linkability is confirmed by showing that our results hold for a set of *actual Twitter users* who tweet from multiple accounts. This has some obvious privacy implications, both positive and negative.

1. INTRODUCTION

Microblogs offer a fast and highly scalable information sharing, allowing multitudes of users to disseminate pithy messages to news-hungry and attention-challenged followers or subscribers. For either social or professional purposes, users share their thoughts, interests and sometimes express highly sensitive or controversial opinions. Twitter, the most prominent microblogging site, has now grown to a truly global service with hundreds of millions of users [2]. In Twitter and similar services, establishing a relationship between users (referred to as tweeters in Twitter) typically requires no reciprocal approval and is often considered as a form of subscription. Although postings (called tweets in Twitter) can be tagged as private, many users make theirs public, as well as their follower information.

At the same time, microblogging has become a rich source of information about individuals. Although it may be difficult to justify the claim that the existence of public messages violates privacy of their authors (since they are the one who make their utterances public in the first place), the potential to de-anonymize multiple user accounts and to link messages (or sets thereof) produced by the same author represents a threat to privacy. This is primarily because multiple accounts owners often expect each set of messages to remain within the boundaries of the account from which it has been posted.

Consider for example, the case of an activist who uses a pseudonym in Twitter[1], to post some highly sensitive or controversial tweets and, in parallel, uses another account associated with his/her real name. Linkage of these two accounts might pose a serious threat to that person's privacy and perhaps even to their physical security.

On the other hand, microblogging site operators and law-enforcement agencies might benefit from techniques that link accounts within a site or across multiple sites. Legitimate reasons might include: (1) identifying spammers and phishers who hide behind multiple accounts to evade detection, (2) tracking or correlating messages pertaining to illegal activities[2], such as terrorism incitement, pedophilia or human/drug trafficking.

Prior work explored linkability of accounts across multiple online services, either relying on similarity of account names, user contributions (including their contents, timestamps, and geo-locations) or account attributes [18, 10, 11]. However, linking multiple accounts within the same online platform, is a challenging task since user attributes and contributions can be deliberately made different or misleading.

[1]Twitter, as opposed to other major OSNs, does not require users to provide real names; pseudonyms are accepted as long as they do not impersonate other users accounts `https://support.twitter.com/entries/18311`

[2]Clearly, the definition of illegal activities varies widely from country to country.

Similarly, a user desiring privacy by using a pseudonym on one account (who also has another account under her real name) would naturally choose a user name unrelated to her actual identity.

In this paper, we investigate the use of probabilistic models to link user accounts in the microblogging ecosystem of Twitter, where message size is limited to 140 characters and where the use of metadata tags (hashtags) and reposts of other users' messages (retweeting) is commonplace.

The goal of this work is to assess linkability of tweets by measuring the extent to which they relate to author's other tweets. We perform our analysis over two large datasets, each containing over $8,000$ Twitter accounts, with over 28 million tweets in one set and over 3 million in the other. The first dataset consists of prolific tweeters, with users producing at least $2,000$ tweets within a six-months period. The second dataset consists of less prolific tweeters who post between 300 and 400 tweets in the same period. Furthermore, we extend our analysis to demonstrate linkability of tweets of actual users who operate multiple Twitter accounts. Our work makes the following contributions:

1. We conduct a large-scale linkability study by applying the Naïve Bayes model as a classifier to predict linkage between tweets using simple stylometric features and analyzing the effect of a specific Twitter feature – hashtags. Results demonstrate that tweets are highly linkable. By simply analyzing letter frequencies, our classifier can – in some scenarios – link tweets with over 90% accuracy. We also show that, by using hashtags alone, our classifier achieves high linkability ratios and can link 2/3 of all tweets.

2. We verify our linkability technique using tweets of real users who actually own multiple accounts. Specifically, we consider a collection of tweets from 14 Twitter users who maintain two or more accounts. By mixing these tweets into a much larger dataset, we successfully classify and link each user's set of tweets from among $8,000$ accounts.

We envision several possible uses of our approach. First, it could be implemented as a service usable by prospective tweeters to assess linkability of their multiple accounts. Second, it can be used by the providers (e.g., Twitter) to identify multiple accounts owners who generate libel, spread disinformation or promote illegal activities.

2. BACKGROUND: NB MODEL

Naïve Bayes (NB) [13] is a probabilistic model based on the so-called Naïve Bayes assumption, which states that all features/tokens are conditionally independent, given some category. In our case, the category is a tweeter (user). Given a document with a set of tokens/features: $token_1, token_2, \ldots, token_n$, NB model computes the corresponding user model as follows:

$$User = argmax_U P(U|token_1, \ldots, token_n)$$

where U varies over all distinct users in our dataset, in order to maximize the probability $P(U|token_1, \ldots, token_n)$ of categorizing a document (represented by $token_1, token_2, \ldots, token_n$) as belonging to a specific user. Using the Bayes Rule[14], $P(U|token_1, \ldots, token_n)$ is defined as:

$$P(U|token_1, token_2, \ldots, token_n) = \frac{P(U)P(token_1, token_2, \ldots, token_n|U)}{P(token_1, \ldots, token_n)}$$

where $P(token_1, token_2, \ldots, token_n|U)$ is the probability of generating the set of tokens $token_1, \ldots, token_n$ by U. When using the Naïve Bayes assumption, $P(token_1, \ldots, token_n|U) = P(token_1|U) \cdots P(token_n|U)$, finding a matching $User$ for $token_1$, $token_2$, $\ldots$, $token_n$ for a given tweeter profile boils down to:

$$User = argmax_U P(token_1|U) \cdots P(token_n|U)$$

where $P(token_i|U)$ is the probability of generating $token_i$ by U. We assume that $P(U)$ – the probability of associating a document to a user without any information about the tokens in the document – is the same for all tweeters. Also, the denominator $P(token_1, \ldots, token_n)$ in the equation above is the combined probability of all tokens; it is therefore independent of the user. Finally, to avoid the under-flow problem, we consider the *log* of the products, which results in:

$$User = argmax_U \sum_i logP(token_i|U)$$

We estimate all probabilities of the form $P(token_i|U)$ using the Maximum Likelihood estimator [7] with Laplace smoothing [14].

3. DATASET

We use a portion of the very large dataset crawled by Yang et al. [22], that spans an approximately six-month period from June to December, 2009, excluding October[3].

The dataset consists of over 400 million tweets, authored by over 15 million users; see Table 1. Majority of users authored at most 10 tweets each, i.e., most are not what we would call prolific tweeters. However, there is still a substantial number of prolific tweeters. For analysis purposes, we extracted two subsets. The first is referred to as Prol. It contains all tweets of all users who authored at least $2,000$ tweets during the observed 6-month period. This set consists of $8,262$ distinct twitter accounts. We consider these to be highly-prolific tweeters. The total number of tweets in Prol is $28,625,352$.

The second subset is referred to as Low. It contains all tweets authored by a set of $10,000$ users, randomly chosen among all users who authored between 300 and 400 tweets. This

[3]Due to some technical difficulties in extracting data for October 2009

Parameter	Value
Total # of Tweets	400,834,808
Total # of Tweeters	15,905,473
Max. # of Tweets per Tweeter	82,177
Min. # of Tweets per Tweeter	1.0
% Tweeters with $\geq$ 2000 Tweets	0.05%
% Tweeters with $\geq$ 500 Tweets	0.68%
% Tweeters with $\geq$ 300 Tweets	1.4%
% Tweeters with $\geq$ 50 Tweets	9.5%
% Tweeters with $\leq$ 10 Tweets	73%
% Tweeters with $\leq$ 1 Tweets	33%

Table 1: Dataset Statistics

corresponds to a total of 73,004 users – a much broader demographic. The total number of tweets in Low is 3,449,635. Since the original dataset is from 2009, given ever-increasing popularity of Twitter, we speculate that Low is a sample of very large demographic in the current state of Twitter.

We extracted two subsets, instead of one, since we aim to assess linkability of tweets for users in different prolificacy scales. We acknowledge that our selection of thresholds in constituting these subsets is subjective. However, we believe that increasing use of Twitter [2] results in more users falling into the ranges of these thresholds. Linkability analysis for users who produce fewer tweets is deferred to future work. Note that we did not filter out re-tweets as they are considered part of users contributions.

4. SETTINGS AND METHODOLOGY

We adopt the settings, linkability-related definitions and abbreviations similar to those in [5]. Our goal is to first assess how much tweets leak about their authors, i.e., how accurately a seemingly anonymous recent tweet can be linked to previous tweets by the same author. Later, in Section 5.5, we analyze how tweets authored by the same tweeter, while using two different accounts, can be re-linked.

We build a Naïve Bayes classifier as our matching/linking model, which is partially trained on two subsets: Prol and Low, to perform linkage. Specifically, for each author U in both Prol and Low, we randomly split her tweets into two sets: Identified Record (IR), and Anonymous Record (AR). The set of all IRs is used for training the classifier and the set of all ARs is used to assess linkability – the accuracy of linking an AR to its corresponding IR. Note that we conduct linkability analysis independently over each set.

For each set and for each author, the task of the classifier is to link her AR to a corresponding IR while maximizing the number of ARs correctly linked. For each set, it matches each author's AR to an IR by returning a list of candidate IRs, sorted in decreasing order of likelihood of being the correct match.

We consider a given AR to have Top-x linkability if the actual corresponding IR is among the top x candidate IR records that the classifier returns. We measure performance of our classifier in terms of linkability ratio (LR), which computes the percentage of ARs that have been correctly classified within the top x candidates. We experimented with three x values: 1, 5, and 10.

We partitioned each author's tweets into IR and AR as follows: First, we sorted a user's tweets in random order. Then, we assigned all tweets (except the last 100) to IR. For the last 100 tweets, we assigned the first y to AR. We vary y over the following values: 5, 10, 20, 50 and 100. The main reason for varying y is to evaluate the impact of the number of anonymous tweets on linkability.

5. LINKABILITY ANALYSIS

We use the Naïve Bayes (NB) model as a matching tool to link tweets based on two types of lexical tokens: (1) unigrams: all letters of the English alphabet, i.e., 26 tokens, and (2) bigrams: all possible two-letter combinations, i.e., 676 tokens. We perform separate analysis on Prol and Low, and compare respective results. For each set, we build a NB model based on all IRs of the set, match all ARs and compute the resulting LR.

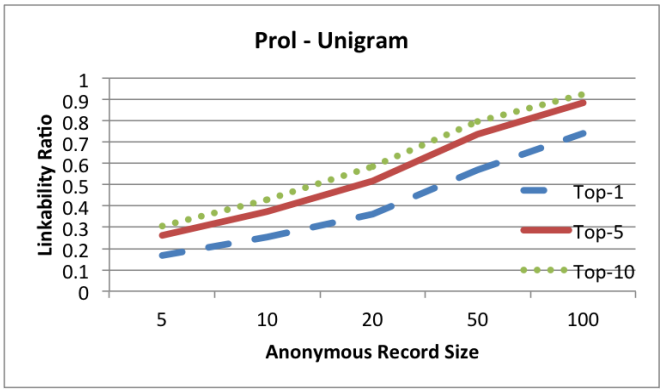
5.1 Unigrams

Figures 1(a) and 1(b) show LRs for Prol and Low, respectively. Specifically, they show Top-1, Top-5 and Top-10 LRs for various AR sizes, using unigrams in NB.

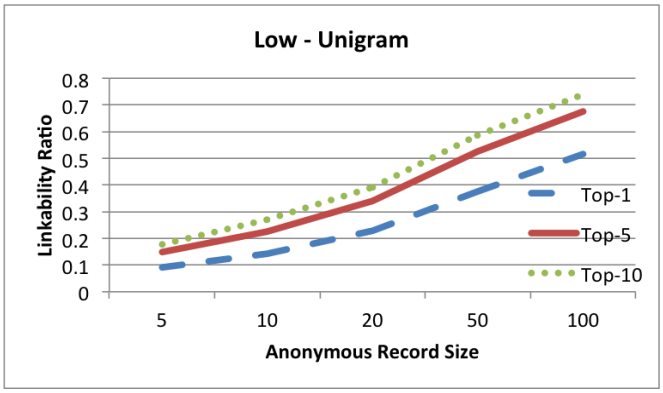
First, we observe that all curves exhibit a clear upward trend, showing that the bigger the AR, the higher LR we obtain. For AR size of 100, Top-10 LR is as high as 92% for Prol and 74% for Low. We also observe relatively high LRs for Top-1(Top-5) – 74% (88%) of ARs are linked in Prol for AR size of 100. Even for small AR sizes, we link a large number of ARs. For example, for AR size of 20, Top-10 LR is 59% for Prol. Even though the number of tweeters is large and the number of tokens is only 26, we can link a substantial number of ARs. As expected, we observe that LRs are higher in sets that have larger IRs (Prol better than Low), since larger IRs offer the classifier more information to capture the user's writing style.

5.2 Bigrams

Figures 2(a) and 2(b) show Top-1, Top-5 and Top-10 LRs, when bigrams are used. Substituting unigrams with bigrams substantially increases LRs, even when using a rather small AR. For instance, Top-1 LR is 95% and 87% for Prol and Low, respectively, for AR size of 100. Even for small AR sizes, we achieve high LRs. For example, for AR size of 5(10), Top-10 LR exceeds 67(84)% in Prol. Also, for AR size of only 20, Top-5 (Top-10) LR in Low is around 75(80)%. As observed earlier with unigrams, sets with larger IRs offer higher LRs, which is again due to offering more information to model user's tweets. Notably, when using bigrams,

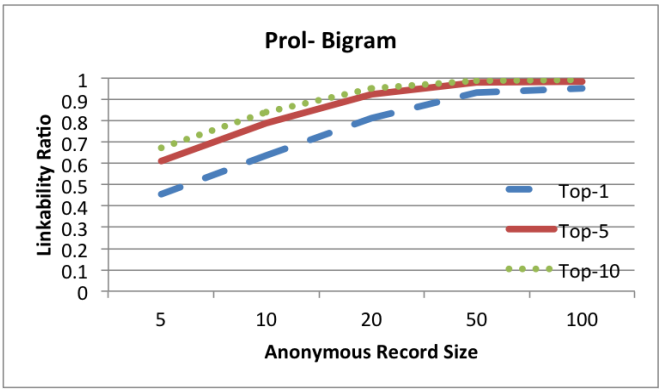


(a)

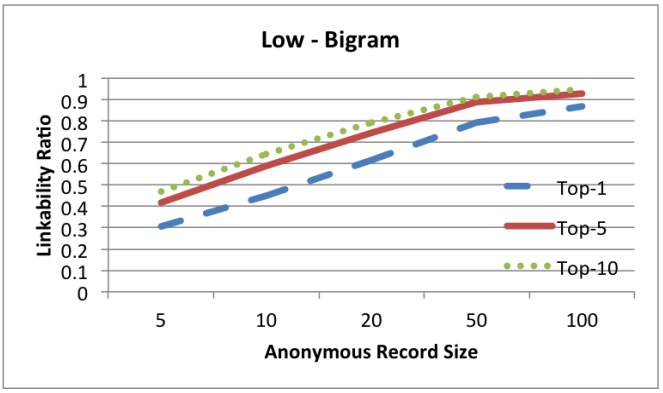


(b)

Figure 1: Top-1, Top-5, and Top-10 LRs of unigram-based NB model for Prol (a) and Low (b)



(a)



(b)

Figure 2: Top-1, Top-5, and Top-10 LRs of bigram-based NB model for Prol (a) and Low (b)

the classifier needs a smaller AR size to achieve substantially higher LRs. For example, using bigrams with an AR size of 20, the classifier obtains a Top-5 LR of over 70% for Low tweeters, while it needs an AR of at least 100 to obtain similar performance with unigrams. These results suggest that there is a tradeoff in choosing between unigrams and bigrams. On one hand, bigrams lead to better LRs with fewer tweets to learn from. On the other hand, unigrams are less computationally demanding and should be considered when there is a need for a better scalability and faster computation.

5.3 Varying the Number of Users

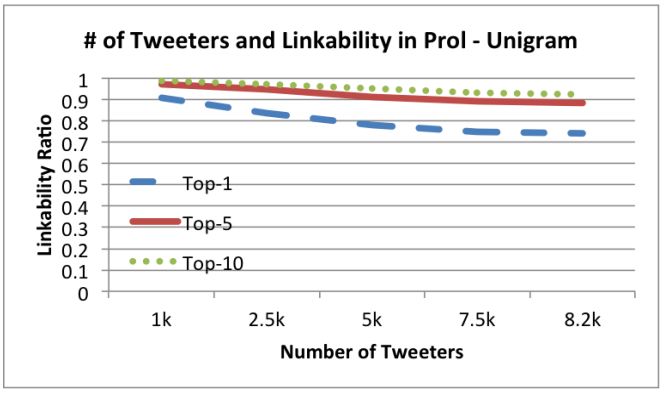
In prior sections, we considered the full set of users, i.e., 8,262 in Prol and 10,000 in Low. We now reduce the number of users in both sets and explore the effects on linkability.

Unigrams. We vary the number of users in Prol and Low between 1,000 and full set size, i.e., 8,262 and 10,000, respectively). Figures 3(a) and 3(b) show LRs for varying set sizes when AR is 100. As expected, LR increases with the smaller number of users. Top-10 LR reaches 99% and 92% in Prol and Low, respectively, when set size is 1,000, which is reasonably large. For set size off 5,000, Top-5 LR exceeds 90% and 70% in Prol and Low, respectively. Note that when we increase to the full set size, reduction of Top-10 LR does not exceed 7% and 20% in Prol and Low, respectively. This points to the resilience of our unigram-based linkability model.

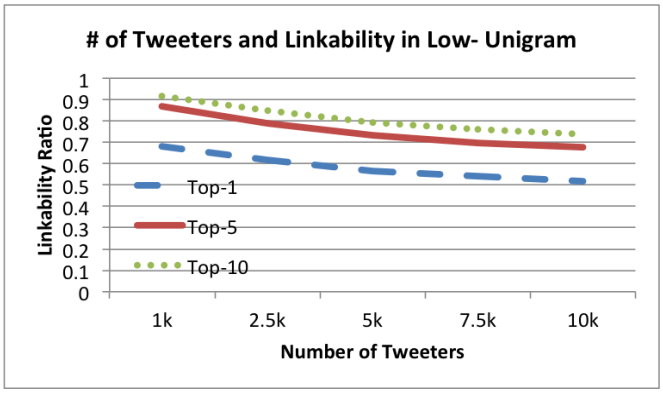
Bigrams. With similar ranges, we vary the number of users in Prol and Low. Figures 4(a) and 4(b) show LRs for varying set sizes with AR size of 100. Interestingly, LR does not decrease much as we increase set size. For example, looking at the entire range, Top-1 LR does not decrease over 4% and 7% in Prol and Low, respectively. Meanwhile, Top-1 LR stays above 95% in Prol and above 87% in Low. Also, in Top-5 and Top-10 LR, the decrease does not exceed 5% in Low and almost 0-1% in Prol. This shows that the bigram-based NB model is very resilient against increasing the number of users. We believe that these results should be very troubling to tweeters worried about linkability of their tweets.

5.4 Improving Unigram-based Model

We previously observed that relying on only bigram tokens yields very high LRs. However, bigrams also require more resources than unigrams: 676 vs 26 tokens.

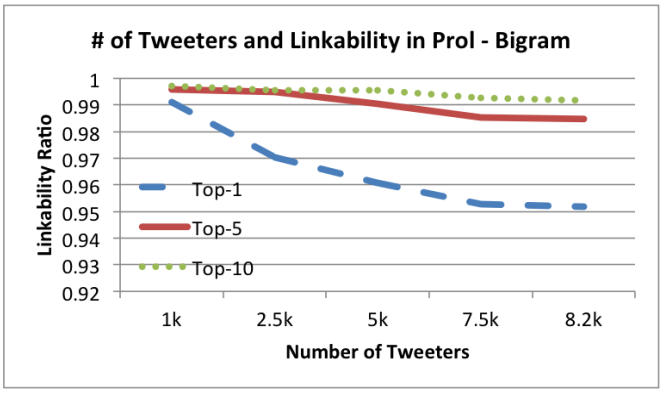


(a)

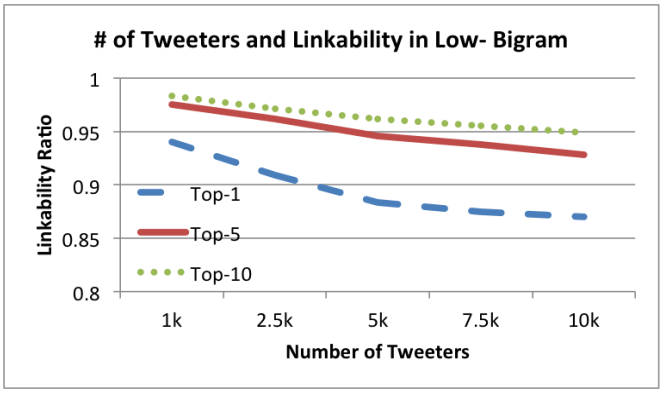


(b)

Figure 3: LRs of unigram-based NB model when varying # users in Prol (a) and Low (b)



(a)



(b)

Figure 4: LRs of bigram-based NB model when varying the number of users in Prol (a) and Low (b)

Thus, bigram-based models are less scalable. To this end, we consider improving LRs when only unigrams are used, by exploring the use of hashtags. We first consider using unigrams from the hashtags themselves, and then combine them with unigrams from full-tweet texts.

Hashtags are a peculiar, yet popular, feature of Twitter. Not surprisingly, many tweets in our dataset contain one or more hashtags. We first filter out from Prol and Low all tweets that do not include any hashtags. We then discard all users with fewer than 300 hashtag-containing tweets; this is so that we can populate their corresponding AR sets with 100 tweets[4]. This leaves us with $3,179$ and 160 users in Prol and Low, respectively. Since the resultant size of Low is small, we confine our analysis to Prol. The number of tweets in filtered Prol is $4,274,188$.

Initially, we intended to use hashtags as tokens in the NB model. However, this yielded a very large number ($> 150,000$) of hashtags, which is very resource-consuming. To remedy the situation, we decided to use unigram tokens within hashtags. Since hashtags can include 11 non-alphabetical characters (i.e., $0-9$ and "_"), we ended up with 37 tokens total.

Figure 5 shows LRs for hashtag-based unigrams in NB. (Recall that our analysis is based on the filtered version of Prol with $3,179$ users.) As can be easily seen, Top 10 LR reaches 67% for AR of 100. Despite only relying on hashtags, we can successfully link 2/3 of ARs.

Combining Hashtags and Full-Tweet Texts.

We consider improving LR by exploring unigrams of full-tweet texts (a set of 26) combined with hashtags-based unigram tokens – a set of 37. Combining these two sets yields 63 tokens, which is still a lot less than 676 with bigrams.

As discussed in Section 2, we use the following $Log - Sum$ to sort the matching users from a set of tokens:

$$\sum_i logP(t_i|U)$$

We now use a weighted average for combining the $Log-Sum$ of full-tweet-text-based tokens and hashtag-based tokens as

[4]We defer to future work the case of linkability when IR size is smaller than AR size.

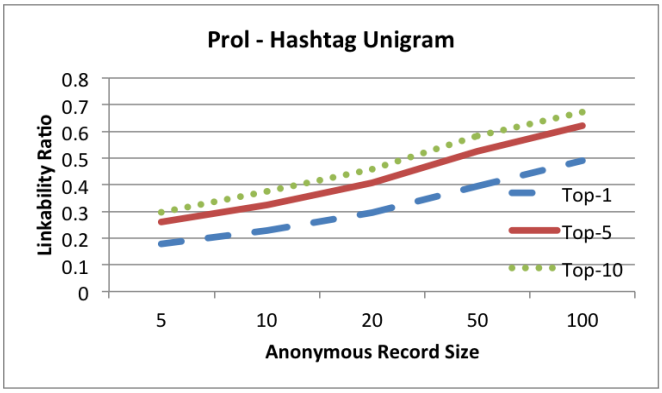


Figure 5: Top-1, Top-5, and Top-10 LRs of hashtag-based NB model when using unigrams

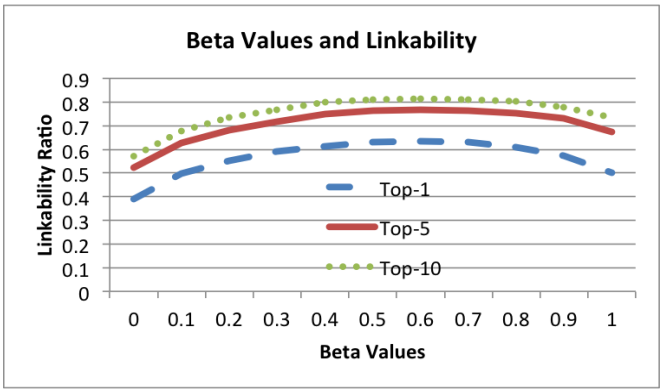


Figure 6: Top-1, Top-5 and Top-10 LRs when varying beta β values from 0 to 1

follows:

$$(\beta) \times \sum_{t_i^{tweet}} logP(t_i^{tweet}|U) + (1-\beta) \times \sum_{t_i^{hashtag}} logP(t_i^{hashtag}|U)$$

However, there is no clear way to assign a value to β. We experimented with several choices. Specifically, we tried all β values ranging in $[0-1]$ at 0.1 increments and observed the highest LR with $\beta = 0.6$ in Top-1, Top-5 and Top-10 LRs. Figure 6 shows LRs for different β values. Note that β selection process (training) is restricted to the set of IRs (ARs are excluded). That is, β is learned using IRs, by further splitting them into identified and anonymized[5] sets, and computing the corresponding LR values in the filtered version of Prol.

Results of Combining. Having chosen $\beta = 0.6$, we used it in computing the weighted average. Figures 7(a), 7(b) and 7(c) show Top-1, Top-5 and Top-10 LRs, when combining unigrams of full-tweet texts and hashtags for the filtered Prol. All figures show LR for β set to 0 (hashtags only), 1 (full-tweet texts only), and 0.6 (combination of tweet texts and hashtags). As evident from the figures, combining full-tweet texts with hashtags substantially boosts LRs for all AR sizes. The improvement over the best of two other curves ranges from $8-13\%$, $6-13\%$ and $4-11\%$ in Top-1, Top-5 and Top-10 LRs, respectively. Note that LRs, when $\beta = 1$, are different from that in Section 5.1. That is because we are assessing LRs of **filtered** Prol, which is much smaller (in terms of number of tweets for IRs) than Prol. We conclude that unigrams in hashtags make tweets linkable, but they are more powerful when combined with unigrams from full-tweet texts. This combination certainly depends on the choice of β.

[5]We allocate the last 50 tweets of IRs as the new ARs used for training.

5.5 Actual Dual-Account Tweeters

So far, we analyzed many sets of tweets, each set authored by a distinct user corresponding to a Twitter account. After *artificially* splitting each such set into Identification Record (IR) and Anonymous Record (AR), we discovered that they are highly linkable. In practice, our technique aims to link **distinct users**, i.e., multiple bodies of tweets emanating from different Twitter accounts. Therefore, results discussed above can be criticized for being too artificial. Indeed, if our approach is truly effective, it must be evaluated using some "ground truth", i.e., real users who tweet via multiple accounts. In this section, we apply the NB classifier to exactly this type of data.

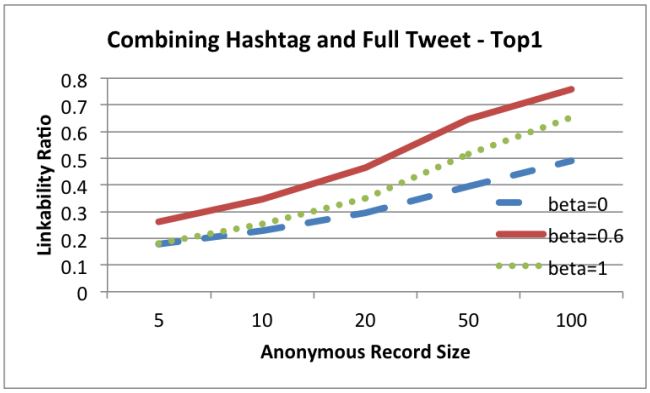
Clearly, we can not assemble a comprehensive collection of all multi-account bodies of tweets, even for a fixed period of time. Instead, as ground truth data, we use a fairly small number of account-pairs that are known to be operated by the same author/user. We mix this new information into the much larger set of tweets used in the previous sections, and then re-apply our NB classifier.

5.5.1 *Multi-Account Dataset*

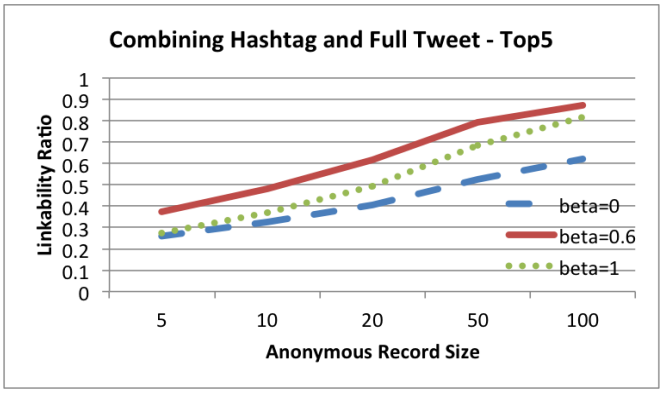
In order to collect tweets from dual-account authors, we manually (and extensively) searched the web for public information pertaining to individuals who operate multiple accounts.. For example, we used search variations of keywords, such as "multiple twitter accounts", and found several blog posts where Twitter users publicly reveal tweeting via multiple accounts. This netted us a total of 41 accounts operated by 14 distinct Twitter users. Of these, 11 actively maintain 2 accounts, while remaining 3 users have 3 or more accounts. For each account, we collect corresponding sets of tweets via Twitter API with twitter-logger[6].

Because of quotas in Twitter API, we could only dump up to ~3200 of a given account's most recent tweets. After crawling 41 accounts, we observed an average number of tweets per account of $1,631$, with a maximum of $3,241$ and a minimum of 34. For each multi-account user, we considered a set of most-prolific two accounts, that have the highest number of tweets. This set constitutes 14 dual-account owners, along with their corresponding tweets, referred as Dual. Most users in Dual operate one relatively general personal

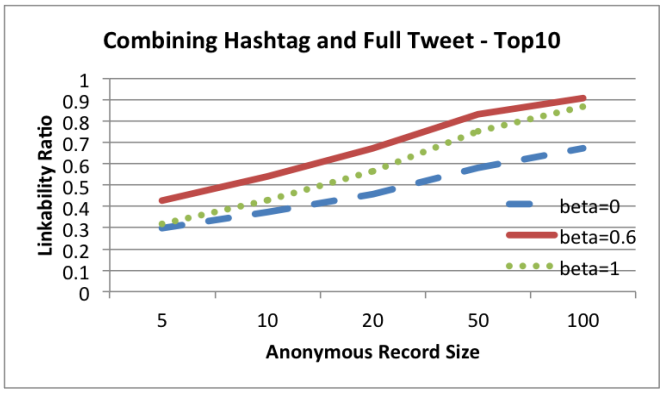
[6]https://dev.twitter.com/, https://github.com/sixohsix/twitter



(a)



(b)



(c)

Figure 7: Top-1 (a) , Top-5 (b) and Top-10 (c) LRs of the revised version of Prol when combining unigrams of full-tweet texts and hashtags

account and another that is more focused on a specific topic, e.g., professional, hobbies, political, or sports.

5.5.2 *Linkability Results*

Based on previously discussed results, we again use bigram NB as the linking model for the Dual dataset. As a sanity check, we first assessed NB's performance with Dual without mixing in tweets from any external datasets. For each dual-account user, we randomly selected one of the two accounts (i.e, all tweets therein) as IR, and the other – as AR.

As before, we varied AR size between 5 to 50, by randomly selecting tweets[7]. Figure 8 shows LRs of 14 users in Dual. Top-1 LR is 100% for AR size of at least 20. Also, Top-1 LR exceeds 85% for AR size less than 10. We therefore conclude that the NB bigram model is very successful in linking tweets from different accounts.

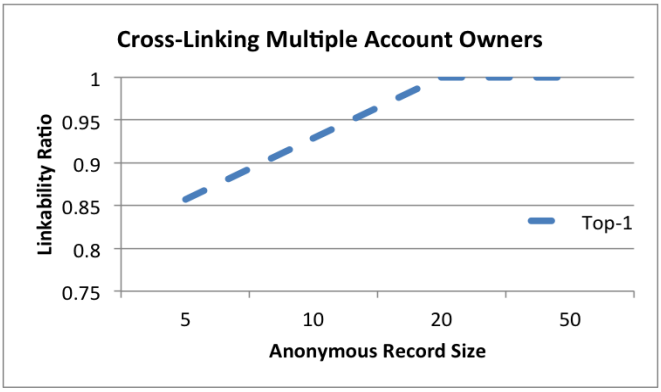


Figure 8: Top-1 LR in Dual – Total number of dual-account owners is 14.

As the next step, we verify that the classifier is scalable for dual-account owners, i.e., performs well if tweets from Dual are merged with Prol and Low datasets, respectively. For this purpose, we merge IRs from Dual with Prol / Low. Likewise, we augment ARs of Dual with ARs of Prol / Low. This leads to $8,262+14 = 8,276$ users in Prol and $10,000+14 = 10,014$ users in Low. Clearly, scaling up the original Dual dataset makes our linkability study really challenging.

Figures 9(a) and 9(b) show LRs of 14 dual-account owners in Dual augmented by Prol and Low, respectively. Once again, Top-1, Top-5 and Top-10 are at 100% when AR size exceeds 20. Surprisingly, for AR size of 10, Top-10 LR exceeds 90% for both cases. This clearly confirms effectiveness of our NB bigram model for the dual-account user linkage when the number of users is large.

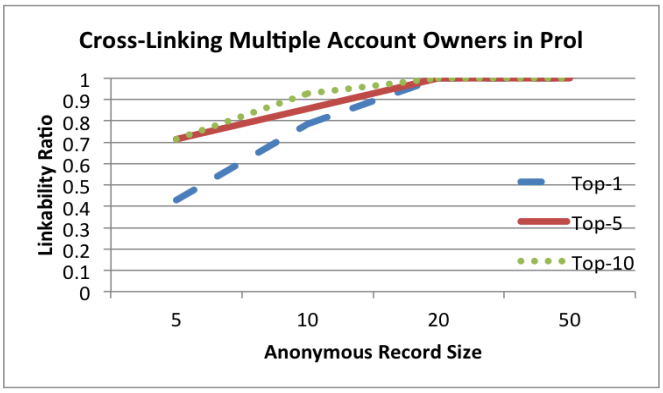
To conclude, our exercise with dual-account owners demonstrates that it is easy to download tweets originated by a very large set of accounts and then effectively link those that are operated by same users. This is clearly detrimental to privacy.

6. KEY RESULTS

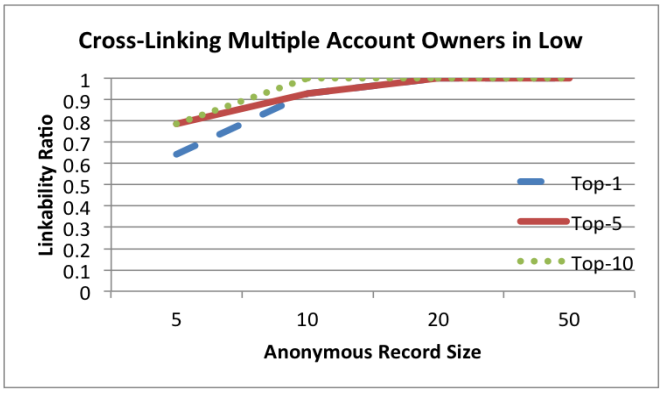
Key elements of our analysis can be summarized as follows:

1. Tweets are highly linkable. Top-1/Top-10 LRs reach up to 95/99% for a large set of users (over $8,000$) by only relying on bigrams; see Section 5.2.

[7]The maximum AR size was set to 50, instead of earlier 100, since some accounts had less than 100 tweets.



(a)



(b)

Figure 9: Top-1, Top-5 and Top-10 LRs in Dual when merging accounts with Prol – (a) and Low – (b)

2. Tweets remain highly linkable even in the context of only unigrams. Top-10 LR exceeds 97% and 85% in Prol and Low, respectively; see Section 5.1.

3. ARs are highly linkable even when their sizes are small. For example, with bigrams and AR size of 20, Top-10 LRs exceed 95% and 79% in Prol and Low; see Section 5.2.

4. LRs are the highest in Prol. However, even with less prolific tweeters, we obtain high LRs. For example, in Low, Top-10 LR reaches 95%, 96% and 98%, for the cases of: 10,000, 5,000 and 1,000 users, respectively (with bigrams); see Section 5.3.

5. Again, with bigrams, LRs do not decrease beyond 7% and 4% in Prol and Low, if we increase the number of users from 1,000 to the full user set size; see Section 5.3. We believe these results are troubling to privacy-conscious tweeters.

6. Unigram-based LR can be substantially improved by combining the unigram model derived from full-tweet texts with that derived from hashtags; see Section 5.4.

7. We evaluated our approach with actual dual-account users merged with Prol / Low. LRs can reach 100% for all 14 dual-account owners that we used as the "ground truth" dataset, as discussed in see Section 5.5.

7. RELATED WORK

Author Attribution in Twitter. Some prior results focused on authorship identification and stylometric analysis of microblogging, e.g., [19, 8, 12, 20, 6]. [19] considered re-identifying authorship of tweets from a set of three authors, while using over 5,000 dimensions as input for a Support Vector Machine (SVM) classifier. Similarly, [12] investigated pseudonymity for a set of 50 Twitter users. [8] studied the use of n-grams with Naïve Bayesian as the linkability model. In particular, 2- to 6-grams are evaluated and a 98% linkability is achieved in the setting of 50 authors. An identification technique based on extracting a set of lexical and syntactical features along with SVM is proposed in [6]. It achieves 91% accuracy for a set of 15 authors. Moreover, a technique based on extracting a set of textual signatures per author is proposed in [20]. Character n-gram 4 and word n-grams are used as features and an accuracy of 30 is achieved when the number of authors is 1,000 (70% accuracy when the number of authors is 50).

There are four main differences between aforementioned results and our work First, we assess linkability on a large scale. The numbers of tweeters is way larger than in prior related studies. Second, we use unigrams, which drastically reduces the number of tokens. Third, we also include hashtags and show their effectiveness in linkability when used alone or in combination with other unigrams. Finally, we successfully re-produce and confirm linkability results for a small set of actual dual-account twitter users.

Cross-Linking Accounts. [10] reports on efforts to cross-link accounts between different social networks, in particular, from Yelp to Twitter and from Flickr to Twitter. Geo-location information, timestamps and text-based features are used in linking user accounts, along with the cosine distance function. There are some notable differences between this study and ours. Techniques that link accounts across social networks might not work for linking accounts within the same network; we do the latter. One reason is that accuracy (across social networks) tends to be high when linking models use similarities among user-names and location information as features (Otherwise, it is quite low). Whereas, a user operating two separate accounts within the same social network would likely not pick similar user-names. Also, location services might be disabled in some social networks.

[4] proposes a technique, based on computing pairwise probabilities, that detects muti-account owners in underground forum (or blog) context. A high precision and recall values are achieved that exceed 90% in certain cases. Other prior work that explored linkability of accounts across platform either relying on similarity of user names or account attributes includes [18] and [11].

Author Attribution. A recent effort [15] investigated de-

anonymization of "anonymous" peer reviews of academic papers. The best result achieved is close to 90%.

One of the best-known results in this area is [3] which uses an extensive set of features, called Writeprints. This approach attained identification accuracy of 91%. Other related studies include [5] which explored author linkability in user reviews using similar (to ours) probabilistic techniques. [16] tackled large-scale online authorship re-identification using linguistic stylometry techniques, achieving up to 80% accuracy. We refer to [21] for an extensive survey of author identification and authorship attribution literature.

De-anonymization in User Preference Databases. Somewhat less related work addressed author identification in the context of movie ratings. Notably, [17] focused on the anonymity of users who rate movies in a sparse Netflix dataset. It defines a model for privacy breaches that relies on an external knowledge base. Then, it shows a de-anonymization attack against a real Netflix dataset [1]. A similar problem is considered in [9] in the setting of public movie discussion forums.

8. DISCUSSION

Although there are many stylometric features in the literature, the focus of this paper is on *scalable* linkability analysis, which motivated us to consider only simple features, allowing the models to perform well with the large number of users.

Even though our technique was evaluated over a very large dataset with numerous accounts, it is possible that some accounts we "linked" are actually operated by a group of authors, rather than by a single one. This might occur if an account serves as an outlet for an organization, e.g., a business, a government agency or a professional society. While privacy concerns are clearly much less serious with such accounts, we believe it is very helpful for their owners to be aware of linkability issues.

Finally, recall that our analysis considered the maximum AR size of 100. While this might seem large, it represents $\leq 5\%$ ($\leq 30\%$) of the total tweets of the user with the minimum contribution in Prol (Low). In practice, it is easy to collect a daily global snapshot of all tweets by using the dedicated Twitter API[8].

9. FUTURE WORK

As part of future work, we plan to improve LRs for smaller AR sizes by considering probabilistic models other than NB (with dependencies among some features) and by looking into other features, such as ratio of tweets to retweets and corse-grained categories of hashtags. We also would like to perform linkability analysis on different author demographics, such as those with very few tweets. Moreover, we believe that further analysis is needed to understand what makes tweets more (or less) linkable. This would allow us to make concrete recommendations for tweeters to retain more privacy.

[8]Accessible via: `https://dev.twitter.com/docs/api/1.1/get/search/tweets`

10. CONCLUSION

This paper reported on a large-scale linkability analysis of tweets. It was based on two datasets, each consisting of over $8,000$ tweeters. We showed that tweets are highly linkable, even when we only rely on unigram and/or bigram distributions. We also performed linkability analysis over a set that included actual dual-account users and linkability remained quite high. Our linkability analysis results should be very troubling for tweeters who are concerned about their privacy.

11. REFERENCES

[1] Netflix. `http://www.netflix.com`.

[2] Twitter Blog. `https://blog.twitter.com/2013/celebrating-twitter7`.

[3] A. Abbasi and H. Chen. Writeprints: A Stylometric Approach to Identity-Level Identification and Similarity Detection in Cyberspace. In *ACM Transactions on Information Systems*, 2008.

[4] S. Afroz, A. Caliskan-Islam, A. Stolerman, R. Greenstadt, and D. McCoy. DoppelgÃd'nger Finder: Taking Stylometry To The Underground. In *IEEE Symposium on Security and Privacy*, 2014.

[5] M. Almishari and G. Tsudik. Exploring linkability in user reviews. In *European Symposium on Research in Computer Security*, 2012.

[6] M. Bhargava, P. Mehndiratta, and K. Asawa. Stylometric Analysis for Authorship Attribution on Twitter. In *Big Data Analytics*, 2013.

[7] C. Bishop. *Pattern Recognition and Machine Learning*. Springer, 2006.

[8] S. R. Boutwell. Authorship Attribution of Short Messages Using Multimodal Features. In *Master Thesis, Naval Postgraduate School*, 2011.

[9] D. Frankowski, D. Cosley, S. Sen, L. Terveen, and J. Riedl. You Are What You Say: Privacy Risks of Public Mentions. In *International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2006.

[10] O. Goga, H. Lei, S. H. K. Parthasarathi, G. Friedland, R. Sommer, and R. Teixeira. Exploiting Innocuous Activity for Correlating Users Across Sites. In *WWW*, 2013.

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

[12] R. Layton, P. Watters, and R. Dazeley. Authorship Attribution for Twitter in 140 Characters or Less. In

Cybercrime and Trustworthy Computing Workshop (CTC), 2010.

[13] D. Lewis. Naive(bayes) at forty:the independence assumption in information retrieval. In *Proceedings of the 10th European Conference on Machine Learning*, 1998.

[14] T. Mitchell. *Machine Learning*. McGraw Hill, 1997.

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

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

[17] A. Narayanan and V. Shmatikov. Robust De-anonymization of Large Sparse Datasets. In *IEEE Symposium on Security and Privacy*, 2009.

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

[19] R. Silva and G. Laboreiro and L. Sarmento and T. Grant and E. Oliveira and B. Maia. Automatic Authorship Analysis of Micro-Blogging Messages. In *NLDB*, 2011.

[20] R. Schwartz, O. Tsur, A. Rappoport, and M. Koppel. Authorship Attribution of Mirco-Messages. In *EMNLP*, 2013.

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

[22] J. Yang and J. Leskovec. Patterns of temporal variation in online media. In *WSDM*, 2011.