De-anonymizing users across heterogeneous social computing platforms

Mohammed Korayem* and **David J. Crandall**
School of Informatics and Computing
Indiana University
Bloomington, IN
* Department of Computer Science, Fayoum University, Egypt
{mkorayem,djcran}@indiana.edu

Abstract

Many people use multiple online and social computing platforms, and choose to share varying amounts of personal information about themselves depending on the context and type of site. For example, people may be willing to share personally-identifiable details (including their real name and date of birth) on a site like Facebook, but may withhold their identity on a dating site that may be widely viewed by strangers. We study the extent to which subtle correlations in a user's activity patterns across different sites may be used to infer that two accounts correspond to the same person. We study a variety of features, including similarity of temporal access patterns, textual content, geo-tags, and social connections, finding that even very weak signals yield surprisingly accurate de-anonymization results.

Introduction

Online social computing platforms have become extremely popular with nearly one billion regular users (Facebook Newsroom 2012). These platforms help people share their status updates, photos, and other content with family and friends. However, with this ease of sharing comes the potential to unintentionally reveal private information. Recent work has shown that sensitive details can be inferred in surprising ways from seemingly innocuous information. For instance, people can be identified based on their web search histories (Barbaro and Zeller 2006) or movie reviews (Narayanan and Shmatikov 2008), Social Security numbers can be inferred from birth date and place (Acquisti and Gross 2009), friendships can be inferred from similarities in travel patterns (Crandall et al. 2010), and so on. These (and other) privacy threats stem from the fact that pieces of information that are uninformative in isolation can become highly distinctive when combined together.

A particular class of privacy threat involves combining information from different social computing platforms. Many people have accounts on multiple websites, and they use these websites in different ways and for different purposes. They may reveal differing types and amounts of information depending on the nature of the site and their trust in the site's privacy and security. For instance, a user may choose to reveal his or her real name and details on Facebook, but may use a pseudonym and claim sanitized (or, perhaps, enhanced) personal details on a dating site. He or she may do this to try to hide their dating activities from their friends, or to hide personal details from amorous strangers. Either way, the implicit assumption here is that it is not possible for others to connect together the person's profiles on the two sites.

Recent work has shown that this is a dangerous assumption. For example, one can use the structural properties of two social graphs to find nodes corresponding to the same individual, even if all other identifying information has been removed (Narayanan and Shmatikov 2009). Other work in digital stylometry (Narayanan et al. 2012) has shown that anonymous authors of online articles and blogs can be identified by analyzing properties of their writing style.

In this paper, we study the extent to which even weaker features of user activity online can be used to de-anonymize users across social computing platforms. Our hypothesis is that correlations in activity patterns, including the temporal patterns of when users are active on different websites, and weak content features, such as similarities in tags and keywords, can be used to successfully find user accounts that correspond to the same person. We evaluate this capability on a dataset consisting of microblog postings on Twitter and tagged images on Flickr, testing the ability of weak temporal, social, textual, and geographic features to identify pairs of accounts that correspond to the same person.

Related work

Privacy on social networking websites has become an important concern as these sites grow at breakneck speeds. Among the active research topics in this area is de-anonymization of social networks, in which the goal is to estimate the identity of online users from data that has (supposedly) been stripped of identifying information. Here we review only the work most directly related to this paper, and refer the reader to several recent survey papers for a more comprehensive literature review (Wu et al. 2010; Ding et al. 2010; Zheleva and Getoor 2011). Perhaps the closest related work in terms of experimental methodology is (Narayanan and Shmatikov 2009), which also collects a dataset from Twitter and Flickr and develops techniques for identifying pairs of accounts across the two sites corresponding to the same user. That work uses the structure of the social network to perform deanonymization, whereas here we use local features including textual, temporal, and geographic properties of content as well as local social connections. Other work has shown that weak information about users, like their group memberships (Wondracek et al. 2010) or tagging behavior (Iofciu et al. 2011) can be used to uniquely identify them.

Digital stylometry (Narayanan and Shmatikov 2008) analyzes text like web pages and blog posts to identify common authorship, although this work typically uses relatively large amounts of text versus the handful of tags that we consider here. Similarities in geo-temporal activities of users have been used to infer social connections between people (Crandall et al. 2010), an idea which we extend to identify distinct user accounts that are operated by the same person.

Methods

We model the de-anonymizing of users on social networks as a binary classification problem: given an account A on one social computing platform and an account B on another platform, we predict whether or not these two accounts correspond to the same individual. We are particularly interested in the case when A and B are on heterogeneous computing platforms, such as when A is on a micro-blogging site and B is on a photo-sharing site. In these cases, we must rely on relatively weak features to make these classification decisions, since for example there is not enough text content to apply techniques of stylometry. In particular, here we consider four types of features:

- *Temporal features:* We hypothesize that if A and B correspond to the same person, the temporal distributions of activity over time are likely to be correlated; e.g. when she is on vacation her activity on both sites might increase, whereas A and B would be silent when she is asleep.
- *Textual features:* We hypothesize that accounts of the same person will use similar tags and other words, reflecting the person's interests and activities.
- *Geographic features:* If A and B correspond to the same person, any geographic information ("geo-tags") on their shared content are likely to be correlated.
- *Social features:* Two accounts corresponding to the same person are likely to have some similar social connections.

We now describe these features in more detail, and then evaluate the discriminative ability of these features on real data.

Temporal activity similarity features

We propose several simple features to measure similarity in the temporal distribution of activities of two user accounts. We simply discretize time into equally-spaced bins, count the frequency of activity in each bin (i.e., number of photos taken or tweets posted), and then generate a high-dimensional vector representing a histogram of activity over the bins. We then apply standard vector similarity measures to produce several features measuring the similarity between the two temporal distributions. In particular, for two temporal vectors u and v corresponding to the activity patterns of two user accounts, we compute a Jaccard similarity score,

$$F_1(u,v) = \frac{\sum_i \min(u_i, v_i)}{\sum_i \max(u_i, v_i)}.$$

We also compute the Hamming similarity between vectors,

$$F_2(u,v) = \sum_i \alpha(u_i, v_i),$$

where $\alpha(\cdot,\cdot)$ is 0 if its parameters are zero and 1 otherwise.

We also compute cosine similarity between vectors,

$$F_3(u,v) = \frac{u \cdot v}{\|u\| \, \|v\|} = \frac{\sum_{i=1} u_i v_i}{\sqrt{\sum_{i=1} u_i^2}\sqrt{\sum_{i=1} v_i^2}}. \quad (1)$$

Before computing cosine similarity, we normalize vectors by dividing each dimension by the total number of users active during that time period. This is the familiar TF-IDF weighting from information retrieval. The intuition in our context is that some periods have much more activity than others (e.g., people take more photos on weekends) so similarities in these popular time periods are less informative.

Finally, we compute a simple statistic that measures the probability that the observed number of matching time bins ($F_2(u,v)$) would result from random chance. Assuming that we select m entries of u at random to be non-zero, and independently select n entries of v to be non-zero at random, the probability that $F_2(u,v) = k$ for $0 \leq k \leq m \leq n \leq N$ is,

$$F_4(u,v) = P(F_2(u,v) = k) = \frac{\binom{m}{k}\binom{N-m}{n-k}}{\binom{N}{n}}, \quad (2)$$

where N is the dimensionality of u and v. Intuitively, this probability is high if the number of matching bins could have arisen by chance, and is low if the degree of similarity is probably due to correlation between the two vectors.

Each of the above similarity functions has advantages and disadvantages, so we compute them all and let a machine learning algorithm (below) decide which are reliable (combinations of) similarity functions for determining if two temporal vectors correspond to the same person.

Text similarity features

Most social platforms allow users to share text, including microblog updates on Twitter or text tags on Flickr photos. Our intuition is that the words a person uses may be similar across different sites and thus reveal identities. As with temporal features, we represent each account with a vector encoding a histogram over words, in the familiar vector space model. We again use TF-IDF weights in which each word count is normalized by the number of people that used it. We compute three similarity values given text vectors u and v: $F_5(u,v)$, the Jaccard distance, $F_6(u,v)$, the cosine similarity between the frequency vectors, and $F_7(u,v)$, the cosine similarity between vectors normalized by IDF.

Geographic similarity features

Modern social media sites allow users to geo-tag content like tweets and photos, and these tags provide another dimension along which to measure similarity between users. A challenge here is that different social media platforms allow geo-tags of different forms and different levels of granularity; for instance, Flickr geo-tags are latitude-longitude coordinates, while Twitter geo-tags are a mixture of coordinates and user-reported place names. To compare these heterogeneous geo-tags, we map latitude-longitude coordinates and user-reported location strings to canonical town names; we used the GeoNames (`www.geonames.org`) database, which includes more than 10 million place names. We then encode the set of geo-tags used by a given user account as a histogram over the town names, producing a vector very similar to that used for measuring text similarity. We use two similarity functions to compare these geo-tag vectors: cosine similarity (F_8), and Jaccard distance (F_9).

We also define a geo-temporal similarity measure that compares the distribution of geo-tags over time. We divide time into equal-length buckets, and compute the geo-tag histogram vector for each of these buckets individually, and then concatenate these vectors together to form a large vector representing the geo-temporal activity of a given user. We then compare these large vectors using cosine similarity (to produce feature F_{10}) and Jaccard distance (F_{11}).

Social connection similarity features

Finally, we define a very simple measure of the similarity of two user accounts' social connections. For Twitter users, we define social connections for a user based on the users mentioned in his tweets and the users he or she re-tweeted, while for Flickr users we collected their public contacts using the Flickr API. Then we simply compare the two sets of social connections, counting the number of user account names in common across the two sets, to produce feature F_{12}.

Classifiers

We then learn classifiers that employ the above similarity features to decide, for an account A on one site and an account B on the other site, whether or not the two accounts are owned by the same person. These algorithms thus learn which combinations of the similarity measures are reliable in this task. We tried three standard classifiers in particular: Decision Trees (Freund and Mason 1999), Naive Bayes (John and Langley 1995; Hall et al. 2009), and linear Support Vector Machines (SVMs) (Chang and Lin 2011).

Experiments and results

To test our ability to connect accounts corresponding to the same person across different social computing platforms, we collected a dataset consisting of photos from Flickr and tweets from Twitter. We used the Twitter API to collect a sample of tweets posted between May 1, 2010 and August 31, 2010, and used the Flickr API to collect a sample photos taken during this same period. We compared the set of Twitter usernames and Flickr usernames across these two datasets, finding 49,585 usernames in common. Of course, accounts with the same username are not guaranteed to correspond to the same person. To reduce this possibility, we examined the hometowns specified on the Flickr and Twitter profiles and removed account pairs for which the two hometown strings did not match (i.e. had Levenshtein edit distance above a threshold). We also removed usernames with missing or very short hometown strings of users. This filtering produced a set of 3,538 people whom we are reasonable sure have accounts on both websites. A manual inspection of a sample of accounts suggests that the error rate is no more than 5% and probably much less (which corroborates the results of (Narayanan and Shmatikov 2009)). Our sample includes 108,206 tweets and 589,045 photos taken by these 3,538 people during the four-month period.

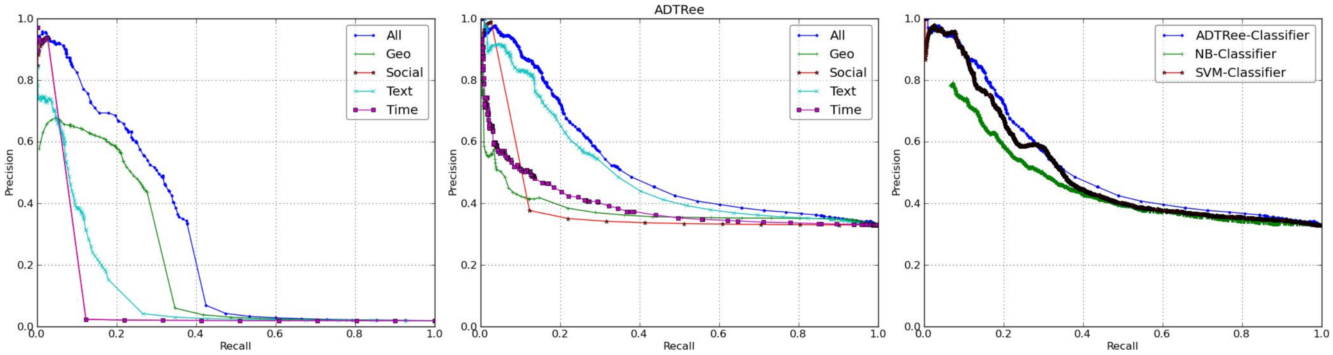
Each of the $\binom{3538}{2}$ possible pairs involving a Twitter account and a Flickr account gives us one exemplar for training or testing, in which the task is to decide if this pair of accounts is operated by the same person. For each of these pairs we computed the 12 similarity features defined in the last section, in addition to several variations on these features. For the four temporal similarity features F_1 through F_4, we computed one set of values using temporal histograms with buckets of size 1 day, and a second set with size 1 hour. We also generated features using both the day and time that Flickr photos were taken, as well as the day and time that photos were uploaded. This yielded 16 temporal features, and a total of 24 similarity features overall.

Having computed feature vectors, we then trained and tested using several classifiers with 10-fold cross validation. In training, an exemplar was considered a positive instance if the pair of accounts were known to correspond to the same person (because the username and hometowns matched) and negative otherwise. During testing, these ground truth labels were hidden from the classifier but used to measure accuracy. We frame this as a retrieval task, in which we wish to find pairs of accounts (one Twitter account and one Flickr account) that correspond to the same person.

Figure 1(a) shows the results of these experiments in terms of a Precision-Recall curve, for different feature types with decision tree classifiers. We observe that the social features are very informative but only for a small subset of users, as evidenced by high precision (about 95%) at only very low values of recall (5%). This is an intuitive result: the set of one's social connections is very distinctive, but useful only for a relatively small set of users who actively use the social features of both websites and whose friends have use the same usernames across the two sites. Time features exhibit a similar pattern. Text features are somewhat more distinctive but relatively weak, e.g. having about 40% precision at 10% recall, while geo-tag features perform best at higher recall rates, and the combination of all features performs better still. For instance, using all features we can de-anonymize about 20% of people with precision of about 68%, or about 40% of people at precision of about 20%.

Figure 1(b) studies an easier task, in which we assume that hometowns are publicly visible in user profiles, so that the set of exemplars consists only of pairs of accounts with similar hometowns. Here there are again 3,538 positive exemplars but only 7,192 negative exemplars. Figure 1(c) compares performance of various classifiers on this easier task, showing that decision trees perform slightly better than SVMs, which perform significantly better than Naive Bayes.

It is important to note the sources of bias that may be present in our datasets. Our technique of linking accounts based on user names and cities may introduce artificial dif-



(a) (b) (c)

Figure 1: Precision-recall curves for retrieving pairs of corresponding users across Twitter and Flickr. *(a):* Performance for different features with the decision tree classifier. *(b):* Performance for different features with decision trees, when negative exemplars are chosen only amongst users in the same hometown. *(c):* Performance of different classifiers using all features.

ficulty, because the ground truth will include some accounts that should be linked (causing our classifier to produce false positives) while matching up accounts corresponding to different people (causing false negatives). On the other hand, users in our dataset may explicitly cross-post to both sites, which would artificially improve the performance of our classifiers compared to users trying to hide their identity.

It is tempting to compare our performance to another recent paper on de-anonymization across Twitter and Flickr (Narayanan and Shmatikov 2009) but this is not easy. Their accuracy measure is weighted by node centrality and ignores singletons, which makes sense for their approach using structural properties of the graph for de-anonymization. In contrast, we measure performance as simply the fraction of users whose privacy we can break versus the number of incorrectly-linked accounts that we would estimate. We plan to compare and combine our approaches in future work.

Conclusion and future work

We examine the extent to which relatively weak features can be used to find accounts on one social computing platform that are owned by the same person on another platform. We take a machine learning approach, developing features based on similarities of temporal, social, textual, and geographic properties of accounts to predict whether pairs of accounts drawn from two sites are owned by the same user. We evaluate this approach on datasets from Flickr and Twitter over a four-month period. Future work could include evaluation on larger datasets consisting of more users over longer periods of time with greater attempt to limit noise and bias, and integration of more sophisticated similarity features.

Acknowledgments

We thank Michael Conover for collecting the Twitter data and for many helpful discussions, Kris Hauser for suggesting the machine learning approach, and Kiran Kumar for the initial exploratory work on this topic.

References

Acquisti, A., and Gross, R. 2009. Predicting Social Security numbers from public data. *PNAS* 106(28):10975–10980.

Barbaro, M., and Zeller, T. 2006. A face is exposed for AOL searcher no. 4417749. *The New York Times*.

Chang, C.-C., and Lin, C.-J. 2011. LIBSVM: A library for support vector machines. *ACM Trans. on Intelligent Sys. and Tech.* 2:1–27.

Crandall, D.; Backstrom, L.; Cosley, D.; Suri, S.; Huttenlocher, D.; and Kleinberg, J. 2010. Inferring social ties from geographic coincidences. *PNAS* 107(52):22436–22441.

Ding, X.; Zhang, L.; Wan, Z.; and Gu, M. 2010. A brief survey on de-anonymization attacks in online social networks. In *Intl. Conf. on Computational Aspects of Social Networks*, 611–615.

Facebook Newsroom. 2012. Key facts. http://newsroom.fb.com/.

Freund, Y., and Mason, L. 1999. The alternating decision tree learning algorithm. In *Intl. Conf. on Machine Learning*, 124–133.

Hall, M.; Frank, E.; Holmes, G.; Pfahringer, B.; Reutemann, P.; and Witten, I. 2009. The weka data mining software: an update. *SIGKDD Explorations Newsletter* 11(1):10–18.

Iofciu, T.; Fankhauser, P.; Abel, F.; and Bischoff, K. 2011. Identifying users across social tagging systems. In *Intl. AAAI Conf. on Weblogs and Social Media*, 522–525.

John, G., and Langley, P. 1995. Estimating continuous distributions in bayesian classifiers. In *Conf. on Uncertainty in Artificial Intelligence*, 338–345.

Narayanan, A., and Shmatikov, V. 2008. Robust de-anonymization of large sparse datasets. In *IEEE Symp. on Security and Privacy*.

Narayanan, A., and Shmatikov, V. 2009. De-anonymizing social networks. In *IEEE Symp. on Security and Privacy*, 173–187.

Narayanan, A.; Paskov, H.; Gong, N.; Bethencourt, J.; Stefanov, E.; Shin, R.; and Song, D. 2012. On the feasibility of internet-scale author identification. In *IEEE Symp. on Security and Privacy*.

Wondracek, G.; Holz, T.; Kirda, E.; and Kruegel, C. 2010. A practical attack to de-anonymize social network users. In *IEEE Symp. on Security and Privacy*, 223–238.

Wu, X.; Ying, X.; Liu, K.; and Chen, L. 2010. A survey of privacy-preservation of graphs and social networks. *Managing and mining graph data* 421–453.

Zheleva, E., and Getoor, L. 2011. Privacy in social networks: A survey. *Social Network Data Analytics* 277–306.