Hybrid De-anonymization across Real-world Heterogeneous Social Networks

Huaxin Li
Shanghai Jiao Tong University
Shanghai, China
lihuaxin003@sjtu.edu.cn

Qingrong Chen
Shanghai Jiao Tong University
Shanghai, China
chenqingrong@sjtu.edu.cn

Haojin Zhu
Shanghai Jiao Tong University
Shanghai, China
zhu-hj@cs.sjtu.edu.cn

Di Ma
University of Michigan-Dearborn
Dearborn, U.S.A
dmadma@umich.edu

ABSTRACT

To enjoy various utility and services, people are active in multiple social networks nowadays. With tons of data generated on platforms, multiple accounts of the same user in different social networks can be used to de-anonymize the user in a large scale. The aggregation of user profiles poses a threat to user privacy. With a concern of privacy leakage, de-anonymization techniques, including graph based approaches and profile based approaches, are widely studied in recent years. However, few works throw light on the de-anonymization between real-world heterogeneous social networks. In this paper, we propose a Hybrid De-anonymization Scheme (HDS) aiming at de-anonymizing heterogeneous social networks. HDS firstly leverages the network graph structure to significantly reduce the size of candidate set, then exploits user profile information to identify the correct mapping users with a high confidence. Performance evaluation on real-world social network datasets shows that HDS has considerable accuracy on de-anonymization and significantly outperforms the prior schemes.

CCS CONCEPTS

•General and reference → General conference proceedings; Evaluation; Experimentation; •Networks → Online social networks; •Security and privacy → Mobile and wireless security;

KEYWORDS

Social Networks Privacy, De-anonymizaion, Heterogeneous Social Networks

1 INTRODUCTION

Along with overwhelming popularity of social networks, people enjoy abundant functionalities and services of a variety of social networks, including sharing status updates, posting photos, communicating with others, and making friends. Due to the different functionalities of different social networks, a user tends to sign in multiple social networks for different purposes. According to the report conducted by Pew Research Center in 2015, 52% of online adults use two or more social media sites such as Facebook, Twitter, MySpace, or LinkedIn[1]. Aggregating user profiles from different social networks reveals various aspects of users. It is interesting that cross-network information represents a double-edged sword. On the one hand, once the user's multiple accounts of different social networks are identified or mapped, these accounts' profiles, preferences, and activities can be collected to benefit personalization, targeting, and recommendation [2, 3]. The latest research pointed out that, the ads delivered by Google, one of the major ad networks, are personalized based on both users' demographic and interest profiles[6]. On the other hand, the adversary can exploit cross-network aggregation to collect the information of various aspects of the target users, which will incur a serious privacy leakage issue [4, 5].

One of the fundamental challenges of bridging the different social identities of the users on different social medias is that the users tend to use varying usernames or have unequal profiles (e.g. fields such as homepage, birthday, etc.) due to the increasing privacy concerns. The process of identifying user from a social network (e.g., anonymized network) based on another social network (e.g., auxiliary network) is called 'de-anonymization'. Recently, there is an increasing interest to study how to 'de-anonymize' or 're-identify' users across social networks, which mainly falls to the following two categories: profile based de-anonymization and structured based de-anonymization, which either suffer from high false positive or assuming the social networks are aligned.

Profile based de-anonymization (or profile matching) exploits the similarities of publicly available profile information of users, such as usernames, text, geographic signatures, and tags to map the users multiple accounts on different social networks [7–9]. Profile matching has the advantages of identifying a specific node of a high confidence, in the case that the accounts belong to the common persons. However, since it takes the whole social network users set as the candidate set, the false positive of profile matching is also high because of the many similar attributes of profiles that belong to different persons in the huge candidate set.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
ACM TUR-C '17, May 12-14, 2017, Shanghai, China
© 2017 ACM. ISBN 978-1-4503-4873-7/17/05…$15.00
DOI: http://dx.doi.org/10.1145/3063955.3063988

Structure based de-anonymization is another widely adopted strategy, which mainly leverages the similarity of social networks' graph structures. In particular, any social network can be modelled as a graph and each user is represented as a node. The relationship between the nodes, such as 'follow' or 'friends', is viewed as an edge. The observation of this kind of approaches is that a user tends to build connections with similar users whom are interested in or acquainted with in different social networks. In other words, these kind of approaches are based on the assumption that the different social networks of the same group users should show the similar network topology, which can be exploited for user identification [10, 11, 13, 14, 16]. The structure based approach provides a promising approach for narrowing down the seed node candidate set. It is effective in the case that two networks are aligned. However, in heterogeneous social networks, this assumption may not always hold due to the fact that the users of different social networks may not be always overlapping. The diversity of usage pattern on different social networks will further render the inconsistency of the network structures of the different social networks. Therefore, in structure based de-anonymization, how to obtain the anchor points and align heterogeneous social networks represents a key challenge.

In this study, we present a Hybrid De-anonymization Scheme for heterogeneous social networks, which is coined as HDS. Different from any previous works which either focus on profile based or structure based approach, HDS aims to integrate the merits of two kinds of approaches. In particular, it firstly leverages the social network structure to significantly reduce the size of node candidate set. Then, it exploits user profile matching to further identify the correct mapping nodes with a high confidence. The seed nodes that act as the anchor points to align two or more heterogeneous social networks will be identified automatically. The major contributions of this paper can be summarized as follows:

- We propose Hybrid De-anonymization Scheme (HDS) to de-anonymize users across heterogeneous social networks. The proposed scheme jointly exploits publicly available network graph structure and user profile information, which is expected to be feasible across real-world heterogeneous social networks and significantly increase the de-anonymization accuracy.
- We conduct extensive experiments on real-world heterogeneous social network datasets to demonstrate the effectiveness of our proposed scheme. The comparative results show that it achieves high detection accuracy and maintains a considerable retrieval rate compared with the direct profile matching.
- Different from most of previous works which mainly de-anonymize anonymized datasets, our study reveals the potential risks to the community of launching de-anonymization attack across real-world social networks, and calls for the following research efforts on privacy-preserving personal recommendation.

The rest of paper is organized as follows. Section II introduces the attacker model and formulates the problem. The proposed approach is presented in Section III. Then, Section IV evaluates the results based on three real-world social networks. Section V discusses related research works, and VI concludes this paper.

2 ATTACK MODEL

We assume two heterogenous social networks G_A and G_U. G_A is denoted as anonymous network and G_U is the auxiliary network. The attacker is able to collect the graph $G = (V, E)$ and profile attributes X_i corresponding to a user $v_i \in V$ by obtaining published datasets or crawling sites. The goal of the attacker is to learn more information of the users across different networks by mapping users in G_A to users in G_U. To achieve this goal, the attacker needs to identify user accounts that belong to a same person in large scale and with a high confidence from two different social networks. This problem can be formally defined as follows.

PROBLEM 1. Given (1) two different social network graphs $G_A = (V_A, E_A)$ and $G_U = (V_U, E_U)$, (2) sets of attributes X_i and X_j of $v_i \in V_A$ and $v_j \in V_U$ respectively, finding user mappings $v_i \leftrightarrow v_j, v_i \in V_A, v_j \in V_U$ that belong to the same real persons accurately by iteratively computing:

$$\underset{v_i \in \text{Cand.}_A, v_j \in \text{Cand.}_B}{\operatorname{argmax}} S(X_i, X_j) \tag{1}$$

where S is a function to compute similarity between X_i and X_j, $Cand._A$ and $Cand._U$ are two candidate sets for potential correct mappings generated by community structure in G_A and G_U, respectively.

Since our proposed de-anonymization approach is based on both graph structure and node (user) profile information, we formally model them respectively as follow.

2.1 Graph Structure Model

Social network structure is usually represented as a graph, where each user is a node in the graph, and connections between a pair of users are represented as edges. Let $G = (V, E)$ represent a social network graph where V is a set of users and $E \subseteq V \times V$, a set of directed/undirected links between users. $e(v_1, v_2)$ means that v_1 and v_2 are in friend relationship or follow relationship where $e \in E, v_1, v_2 \in V$. As an important structure in the social network graph, *community* is formally defined as follows:

DEFINITION 1. A community C in a social network graph is a disjoint partition, which corresponds to a social circle where nodes are closely connected in $G(V, E)$. We denote communities in a graph as $\mathcal{C} = \{C_1, C_2, ..., C_k\}$, where $C_i \neq \emptyset$ and $C_i \cap C_j = \emptyset$ if $i \neq j$ for $1 \leq i, j \leq k$. $\forall C_i \in \mathcal{C}, V_{C_i} \subset V$ and $E_{C_i} \subset V_{C_i} \times V_{C_i}$.

2.2 Profile Information Model

Social networks must allow part of user profile information to be available to public. To exploit profile information across heterogeneous social networks, we firstly give a uniform definition:

DEFINITION 2. Let $X_i = [x_{ik}]_{k=1...d}$ denote a set of attributes associated with the user $v_i \in V$ (for instance, username, location, self-description, etc), where d is the number of types of attributes and x_{ik} records the content of the kth attribute of user v_i. If a user v_i's jth attribute is not available on the social network (e.g., Tom chooses not to show his hometown on Twitter), then $x_{ij} = null$.

Since heterogeneous social network platforms contain different kinds of profile information, and some of which contains semantic or syntactic meaning, mapping two users' accounts from two

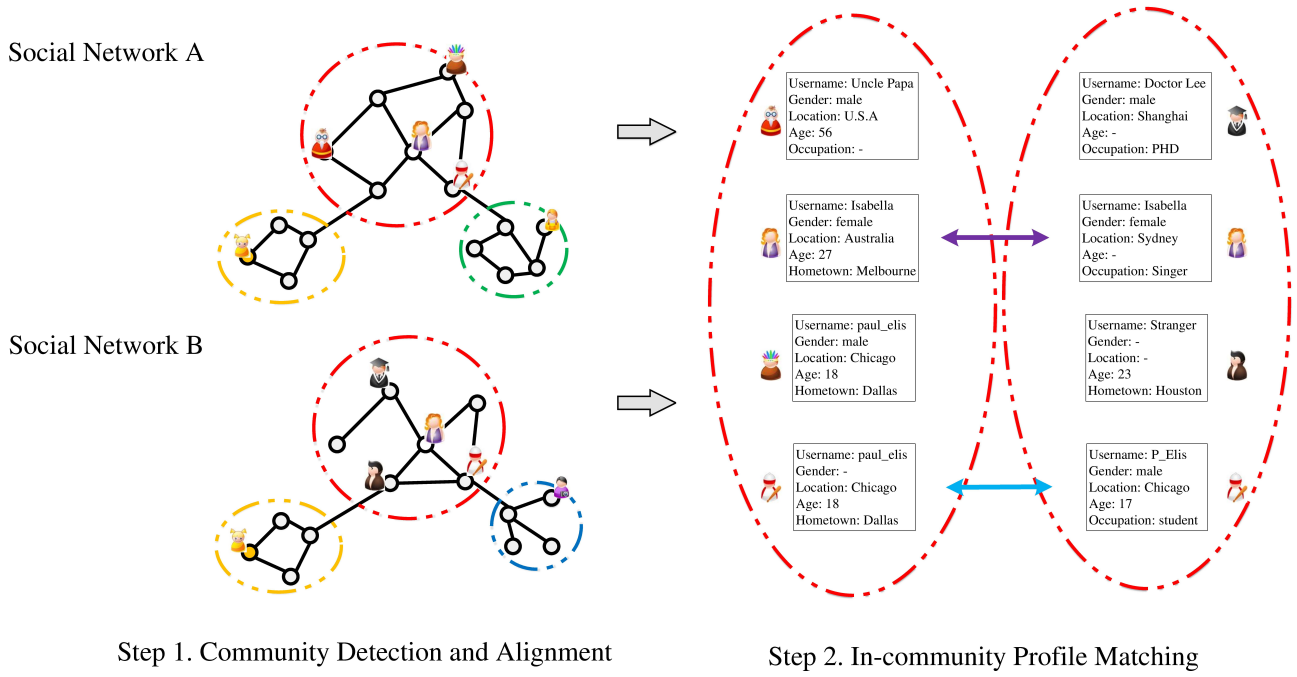


Figure 1: Overview of our scheme

heterogeneous social networks is similar to an ontology matching problem. In general, ontology matching determines an alignment for a pair of ontologies O_1 and O_2. Each ontology consists of a set of discrete attributes which are usually represented in the form of tables, classes, properties, and determines as output the relations. In our problem, profile matching can be defined as follows:

DEFINITION 3. Given two profiles, $p_A = \{x_1, ..., x_A\}$ and $p_U = \{x_1, ..., x_U\}$. If $type(x_i) = type(x_j)$ for two attributes $x_i \in p_A$ and $x_j \in p_U$, the similarity between the two attributes is defined:

$$sim_a = matchScore(x_i, x_j) \tag{2}$$

Then the similarity of profiles is computed by:

$$sim_p = \frac{\sum_{r=1}^{t} w_r (sim_a)_r}{t} \tag{3}$$

where w_r is the weight given to attributes, and t is the number of attribute pairs of the same type between two profiles.

3 HYBRID DE-ANONYMIZATION SCHEME

In this section, we present our proposed Hybrid De-anonymization Scheme (HDS).

3.1 Scheme Overview

Figure. 1 illustrates our proposed scheme which has two main steps: (1) Communities Detection and Alignment: communities in both networks are detected according to graph structure, and communities that contain the same pairs of seeds are aligned, (2) In-community profile matching: in each pair of aligned communities, nodes with high similarity score, which is computed by profile similarity, is accepted as a mapping. Algorithm 1 presents the whole procedure, and the details and time complexity are introduced in the following sub-sections.

3.2 Communities Detection and Alignment

The goal of first step is to partition social network graphs G_A and G_U into two sets of communities $\mathcal{C}_A = \{c_1, ..., c_m\}$ and $\mathcal{C}_B = \{c_1, ..., c_n\}$. We apply Infomap algorithm [28], which has a low time complexity, to partition disjoint, non-overlapping communities $\mathcal{C}_A$ and $\mathcal{C}_U$ for two graphs, respectively. In brief, Infomap finds the shortest multilevel description of the random walker therefore giving us the best hierarchical clustering of the network - the optimal number of levels and modular partition at each level - with respect to the dynamics on the network. So another merit of using Infomap algorithm is that it generates $\mathcal{C}_A, \mathcal{C}_U$ with different scales at different levels so that we can choose communities with similar scale for aligning. The algorithm for communities detection and division is denoted as the $CommDetection(\cdot)$ function in Algorithm 1, and the time complexity is $O(|E|)$.

For aligning communities $C_i \in \mathcal{C}_A$ and $C_j \in \mathcal{C}_U$, [14] proposes to treat each community as a node in a graph, then propagate the communities mapping process from some 'community seeds' using an improved version of [13]. However, in practice, we find that communities can be more easily aligned given the publicly available profile information. As shown in [9, 13], the possibility that two accounts with same usernames do not belong to a user is less that 5%. Thus we align $\mathcal{C}_A$ and $\mathcal{C}_U$ according to the number of same

Algorithm 1 Algorithm of proposed scheme

Input: $G_A < V_A, E_A >, G_U < V_U, E_U >$, threshold θ
Output: Mappings of users μ'

//Communities detection and alignment
$\mathcal{C}_A =$`CommDetection` (G_A)
$\mathcal{C}_U =$`CommDetection` (G_U)
$\mu =$`SelectSeeds` (V_A, V_B)
$CommPairs =$`AlignCommunities` $(\mathcal{C}_A, \mathcal{C}_U, \mu)$

//In-community profile matching
$\mu' =$`InCommunityMapping` $(CommPairs, \theta)$
output $\mathcal{A}$

procedure INCOMMUNITYMAPPING($CommPairs, \theta$)
  $\mu' = \emptyset$
  for $< C_i, C_j > \in CommPairs$ **do**
    for $u_i \in C_i$ **do**
      for $u_j \in C_j$ **do**
        if `MongeElkan` $(u_i, u_j) > \theta$ **then**
          add u_i, u_j into μ'
        end
      end
    end
  end
  Return μ'
end procedure

usernames in communities according to the algorithm described as the following two steps.

The first step is to find all user pairs with same usernames $\mu = \{..., (u_i, u_j)_k, ...\}$ where $u_i \in V_A$ and $u_j \in V_U$. Greedy searching will cause a high complexity of $O(|V_A||V_U|)$. Instead, this process can be implemented by a hash table so that the time complexity can be reduced to $O(|V_A| + |V_U|)$. This procedure is denoted as $SelectSeeds(\cdot)$ function in Algorithm 1.

In the second step, an initial confidence score $cs_{i,j}$ (that indicates whether two communities should be aligned) for each pair of communities (C_i, C_j), where $C_i \in \mathcal{C}_A, 1 \leq i \leq m, C_j \in \mathcal{C}_U, 1 \leq j \leq n$, is set as 0. For each pair $(u_p, u_q) \in \mu$, $cs_{i,j}$ is added by one, given $u_p \in C_i$ and $u_q \in C_j$. Then, all confidence scores cs for communities pairs is examined, if $cs_{i,j}$ exceeds a threshold θ_{cs}, C_i and C_j are aligned. The time complexity of this step is $O(|\mu|) < O(|V_A| + |V_U|)$. This procedure is denoted as $AlignCommunities(\cdot)$ function in Algorithm 1.

The overall complexity of communities alignment algorithm is $O(|V_A| + |V_U|)$, as described above. Our overall evaluations show that our communities division and alignment only slightly reduce the recall rate.

3.3 In-community Profile Matching

$InCommunityMapping(\cdot)$ function in Algorithm 1 describes our in-community profile matching algorithm. Within each pair of aligned communities, users' profiles are pair-wise compared and a similarity score is calculated. If the score exceeds a pre-defined threshold θ, u_i and u_j are accepted as a successful mapping. Due to the social networks' settings and users' willing for sharing information, not all profile attributes are available for all users across multiple social networks. So in this paper, we only select username as the profile information. Syntactic matching is applied to attributes that are usually shown as strings (e.g. username and person name). These attributes on different social networks often have editing differences, such as difference among "Jones, David", "David Jones", and "D. Jones". So string matching metric can be used for syntactically matching these attributes. In order to avoid the influence of abbreviation or acronym, Monge-Elkan algorithm, a recursive string matching algorithm, is applied [29]. The basic idea of this method is to break input string into tokens. Then the best matching token are compared to get the score as follows.

$$MongeElkan(A, B) = \frac{1}{|A|} \sum_{|A|}^{i=1} max\{dist(A_i, B_j)\}_{j=1}^{|B|} \quad (4)$$

where A and B are two strings, and $dist()$ refers to a secondary distance function used to compute similarity between tokens of A and B. In a lot of functions computing edit-distance like functions, *Jaro-Winkler similarity* is chosen as the secondary distance function in our problem, due to its noticeable performance in previous research on name-matching tasks [30]. Monge-Elkan algorithm returns 1 if two string are fully matched or one abbreviates the other; returns 0 if there is no match between the two strings.

4 EVALUATIONS

In this section, We evaluate our proposed HDS scheme by conducting experiments on a set of real-world social networks data.

4.1 Datasets

The datasets of three real-world heterogeneous online social networks, i.e., Last.fm, Livejournal, and Myspace, are obtained from [23]. The datasets include node information, edge information, and profile information of a subset of users of these social networks. We evaluate our proposed scheme on the three social networks pairwise.

- *Last.fm* is the world's largest online music catalogue and has been recognized as a popular social network for music enthusiasts. Last.fm builds detailed profiles of users musical tastes and preferences. The dataset consists of 136,420 users and 1,685,524 friend relationship.
- *LiveJournal* is a social networking site and blogging platform that allows users to find each other through journaling and interest-based communities. The dataset consists of 3,017,286 users and 19,360,690 friend relationship.
- *MySpace* is a social networking website offering an interactive, user-submitted network of friends, personal profiles, blogs, groups, photos, music, and videos. The dataset consists of 854,498 individuals and 6,489,736 friend relationship.

We build undirected social network graphs according to 'friend' or 'follow' relationship in these social networks. The statistics of the graphs are shown in Table 1. In order to evaluate the results, we

obtain the ground truth data from [23, 31], which contain pair-wise matched user id of two social networks. The data were originally collected by Perito el. al [31] through Google Profiles service by allowing users to integrate different social network services.

Table 1: Statistics of social networks

Network	Nodes	Edges	Av. Degree
Last.fm	136,420	1,685,524	24.71
Livejournal	3,017,286	19,360,690	12.83
Myspace	854,498	6,489,736	15.19

4.2 Performance of Proposed Scheme

To quantitatively evaluate the algorithm, we consider the two widely-used metrics:

- *Accuracy:* In all mappings returned by the de-anonymization algorithm, the percentage of correct mapping. Since our goal is to find out correct mappings but not to find out incorrect mappings, the concept of accuracy here is same as the concept of precision.
- *Retrieval:* Retrieval, or retrieval rate, is defined as the percentage of correct mapping retrieved by algorithm in all mappings in ground truth.

Table 2 shows how our HDS scheme performs by tuning threshold θ. As introduced in Section 3.3, the threshold θ is a similarity criterion that accepts a pair of nodes as a mapping in our algorithm, i.e. if the similarity score between two nodes exceed the θ, the two nodes are accepted as a potential mapping. So, the higher θ is set, the more similar the accepted nodes are, and thus fewer potential mappings will be returned. So θ actually reflects the trade-off between accuracy and retrieval. An attacker can choose θ according to his/her requirement of this trade-off in practice. When the threshold is set to 0.9, the accuracy of matching users can be more than 90% with a recall of 30% for Livejournal-Lastfm and Livejournal-Myspace. The results show that large scale accurate de-anonymization across real-world heterogeneous social networks can be performed.

4.3 Comparing with Existing Works

Since our scheme is a combination of graph structure and publicly available profile information, we evaluate the results by comparing our approach with approaches that only exploit profile information, and approaches that are only based on graph structure, respectively.

State-of-the-art graph-based de-anonymizing algorithms have been discussed and compared in [19]. However, only a few of them are suitable to real-world heterogeneous social networks for various reasons. Some techniques are constrained by their restricted requirements of the same size of social networks (or same number of nodes) [10, 11], sybil users [12] or high computation capability for large scale networks [17], while others have only been evaluated between the original graph and the noisy graph [14, 16, 20]. For reference, we test the well-known graph-based NS algorithm proposed by Narayanan and Shmatikov [13] and percolation-based de-anonymization algorithm [32] by using the open-source evaluation system proposed in [19]. As a result, only few correct mappings are reported by the two algorithms on the heterogenous social network datasets by feeding more than 100 seeds. One possible reason of the results of graph-based approaches is that pure graph-based approaches require enough overlaps of the network graphs to propagate and correct false mappings at the beginning of the mapping [13]. But it is usually difficult to obtain datasets with ideal overlaps from two heterogenous social networks, which limit the performance of graph-based approaches in practice. According to [13], 30.8% of the mappings were re-identified correctly between a Twitter dataset (Av. degree of 37.7) and a Flickr dataset (Av. degree of 32.2), which is far away from our results on more heterogenous networks datasets. The results show that introducing profile attributes of nodes obviously increases the successful rate of de-anonymizing.

On the other hand, previous studies exploit various user profile information to connect individuals between social networks, including usernames [8, 22], tags [7], activities [21], and multiple kinds of profile attributes [9]. To compare the performance of our HDS, the direct profile-based matching represented by [8, 22], i.e. computing profile similarity between each user in one social network and all users in the other social network and find the most similar one, is used as the baseline.

Figure 2(a) illustrates that the accuracy of our HDS obviously outperforms direct profile matching while the retrieval is just slightly reduced, when the θ is set as 0.9, which is the medium threshold in Table 2. The accuracy of our approach achieves 86%, 95%, 96%, which outperform the direct profile matching by 9%, 7%, 12%, respectively. And the decrease of retrieval rate is less than 4%. And we can also get similar conclusions when tuning the θ. The results reflect that graph structures (community) are useful to filter out incorrect matchings, thus increasing the accuracy.

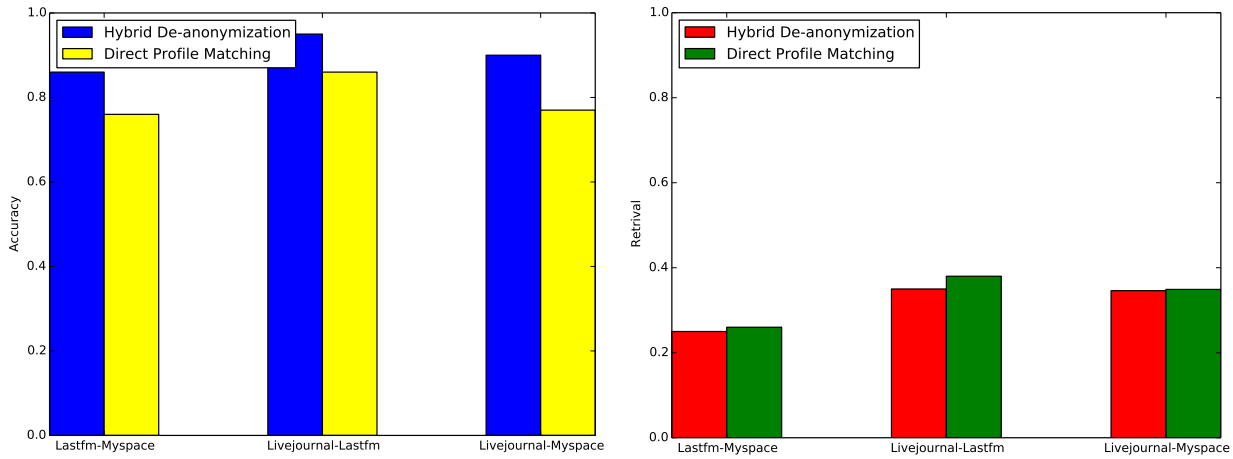
5 RELATED WORK

5.1 Structure based de-anonymization

De-anonymizing social networks is a hot research topic in recent years. Structure based de-anonymization works are based on the assumption that the different social networks of the same group users should show the similar network topology, which can be exploited for user identification [10, 11]. The observation of this kind of approaches is that a user tends to build connections with similar users they are interested in or acquainted with in different social networks. Backstrom et al. introduced both active attacks and passive attacks to de-anonymize social data [12]. Narayanan and Shmatikov performed the de-anonymization attack to large-scale directed social networks. They designed a de-anonymization algorithm by identifying some seeds and propagating based on structure similarity[13]. In [14], Nilizadeh et al. extended Narayanan and Shmatikov's attack by proposing a community-enhanced de-anonymizing scheme of social networks. Then, Lai [15] proposed to detect communities in social networks via user's interests and de-anonymize users in communities. Ji et al. also designed an Adaptive De-Anonymization framework for the scenario that the anonymized and auxiliary graphs have partial overlap [16]. Srivatsa

Table 2: De-anonymization Performance

	θ	0.85	0.86	0.87	0.88	0.89	0.90	0.91	0.92	0.93	0.94	0.95
Lastfm-Myspace	Accuracy	0.61	0.67	0.73	0.76	0.82	0.86	0.89	0.92	0.94	0.95	0.95
	Retrieval	0.28	0.27	0.27	0.26	0.25	0.25	0.24	0.24	0.23	0.23	0.23
Livejournal-Lastfm	Accuracy	0.70	0.77	0.82	0.88	0.90	0.95	0.97	0.98	0.98	0.98	0.98
	Retrieval	0.40	0.39	0.37	0.36	0.35	0.35	0.34	0.33	0.33	0.32	0.31
Livejournal-Myspace	Accuracy	0.69	0.75	0.82	0.85	0.88	0.90	0.91	0.92	0.93	0.94	0.94
	Retrieval	0.38	0.37	0.36	0.36	0.35	0.34	0.33	0.32	0.32	0.32	0.30



(a) Accuracy (b) Retrival

Figure 2: Comparisons with direct profile matching when $\theta = 0.9$

and Hicks modeled mobility traces as contact graphs and presented different attacks for de-anonymizing using online social networks as side channel [17]. However, in heterogeneous social networks, this assumption may not always hold due to the fact that the users of different social networks may not always be overlapping. The diversity of usage pattern on different social networks will further render the inconsistency of the network structures of the different social networks. In our proposed method, we also exploit semantic publicly available information, such as user profile, to help de-anonymize users.

Besides, Ji et al. [18] conducted the comprehensive quantification on the de-anonymizability of 24 real-world social networks with seed information in general scenarios. Later, in [19], a uniform and open-source secure graph data sharing/publishing system was proposed. Qian et al. leveraged background knowledge graph to improve the de-anonymization performance [20]. But this work mainly focuses on de-anonymizing a graph anonymized from original graph and inferring some private attributes. In our work, we try to de-anonymize heterogeneous social networks by considering both semantic information and structure information.

5.2 Profile based user matching

Public information and semantic information on social media or social network sites provide the evidence to match users of different social networks. Iofciu et al. used tags to identify users across social tagging systems such as Delicious, StumbleUpon and Flickr [7]. Goga et al. identified accounts on different social network sites that all belong to the same user by exploiting only innocuous activity, such as location profiles, timing profiles, language profiles, that inherently comes with posted content [21]. In [8], Zafarani et al. matched accounts according to usernames among 12 different Social Web systems. The recent work by Zafarani et al. [22] conducted an in-depth investigation of this problem by defining sophisticated features to model the behavior patterns of users in selecting usernames. Korayem et al. extracted four kinds of features, i.e. temporal activity similarity features, text similarity features, geographic similarity features, social connection similarity features, and apply machine learning techniques to find correct mapping [9]. Zhang et al. [23] connected social networks users by considering both local and global consistency among multiple networks, but they treat both two consistencies as features and train an energy-based learning model. In [24] and [25], the first privacy-preserving personal profile matching schemes for mobile social networks was proposed by Li et al. In this scheme, an initiating user can find from a group of users the one whose profile best matches with his/her, with limited risk of privacy exposure. Later, novel fine-grained private profile matching protocols were designed in [26, 27]. Different from these works, our proposed approach uses social structure to narrow down the candidate sets in order to achieve higher accuracy.

6 CONCLUSION

In this paper, we propose a practical Hybrid De-anonymization Scheme (HDS) for de-anonymizing real-world heterogeneous social networks. HDS is a de-anonymizing scheme that exploits the network graph structure to significantly reduce the size of candidate set, and uses user profile information to identify users with a high confidence. The performance evaluations of our proposed scheme based on a dataset of three real-world social networks show that it achieves high accuracy with a slight sacrifice of retrieval rate. And the comparisons with current works show that our proposed scheme is effective to de-anonymize real-world heterogeneous social networks. Nodes anonymization and privacy preserving in social networks should become more serious concernss.

ACKNOWLEDGMENTS

This word is supported by National High-Tech R&D (863) Program (no. 2015AA01A707).

REFERENCES

[1] M. Duggan, N. Ellison, C. Lampe, A. Lenhart and M. Madden, "Social Media Site Usage 2014, Pew Research Center", 2015. http://www.pewinternet.org/2015/01/09/social-media-update-2014/.

[2] D. Lazer, A. Pentland, L. Adamic, S. Aral, A.-L. Barabsi, D. Brewer, N. Christakis, N. Contractor, J. Fowler, M. Gutmann, T. Jebara, G. King, M. Macy, D. Roy, and M. Van Alstyne, "Computational social science", *Science*, 323(5915):721C723, 2009.

[3] W Gauvin, C. Chen, X. Fu, B. Liu, "Classification of Commercial and Personal Profiles on MySpace", IEEE *SESOC*, 2011

[4] H. Li, H. Zhu, S. Du, X. Liang, and X. Shen, "Privacy leakage of location sharing in mobile social networks: Attacks and defense", In *IEEE Transactions on Dependable and Secure Computing* vol: PP, Issue: 99, pp: 1-1, 2016.

[5] H. Li, Z. Xu, H. Zhu, D. Ma, S. Li, K. Xing, "Demographics inference through Wi-Fi network traffic analysis," In IEEE INFOCOM 2016-The 35th Annual IEEE International Conference on Computer Communications (pp. 1-9), IEEE, 2016.

[6] W. Meng, R. Ding, S. P. Chung, S. Han, and W. Lee, "The Price of Free: Privacy Leakage in Personalized Mobile In-App Ads", *NDSS*, 2016.

[7] T. Iofciu, P. Fankhauser, F. Abel, and K. Bischoff, "Identifying Users Across Social Tagging Systems", *ICWSM'11*, 2011.

[8] R. Zafarani and H. Liu, "Connecting Corresponding Identities across Communities", *ICWSM'09*, 2009.

[9] M. Korayem and D. Crandall, "De-Anonymizing Users Across Heterogeneous Social Computing Platforms", *ICWSM'13*, 2013.

[10] N. Korula and S. Lattanzi. "An efficient reconciliation algorithm for social networks", *Proceedings of the VLDB Endowment*, 7(5), 377-388, 2014.

[11] P. Pedarsani, D. R. Figueiredo, and M. Grossglauser. "A bayesian method for matching two similar graphs without seeds". *Allerton*, 2013.

[12] L. Backstrom, C. Dwork and J. Kleinberg, "Wherefore art thou r3579x?: anonymized social networks, hidden patterns, and structural steganography", ACM *WWW'07*, 181–190, 2007.

[13] Narayanan A, Shmatikov V, "De-anonymizing social networks", In *30th IEEE Symposium on Security and Privacy*, (pp. 173-187), 2009.

[14] Nilizadeh S, Kapadia A, Ahn Y Y. "Community-enhanced de-anonymization of online social networks", In *Proceedings of the 2014 ACM SIGSAC Conference on Computer and Communications Security*,(pp. 537-548), 2014.

[15] S. Lai, H. Li, H. Zhu, N. Ruan, "De-anonymizing Social Networks: Using User Interest as a side-channel", In *2015 IEEE/CIC International Conference on Communications in China (ICCC)*, (pp. 1-5), 2015.

[16] S. Ji, W. Li, M. Srivatsa, J. He and R. Beyah, "Structure based Data Deanonymization of Social Networks and Mobility Traces", Springer *Information Security*, 237–254, 2014.

[17] M. Srivatsa and M. Hicks. "Deanonymizing mobility traces: Using social networks as a side-channel", In *Proceedings of the 2012 ACM conference on Computer and communications security*, (pp. 628-637), 2012.

[18] S. Ji, W. Li, N. Gong, P. Mittal, and R. Beyah. "On your social network de-anonymizablity: Quantification and large scale evaluation with seed knowledge",*NDSS*, 2015.

[19] S. Ji, W. Li, P. Mittal, X. Hu, and R. Beyah, "SecGraph: A Uniform and Open-source Evaluation System for Graph Data Anonymization and De-anonymization", In *USENIX Security*, (pp. 303-318), 2015.

[20] J. Qian, X.-Y. Li, C. Zhang, and L. Chen, "De-anonymizing Social Networks and Inferring Private Attributes Using Knowledge Graphs", In *The 35th Annual IEEE International Conference on Computer Communications (INFOCOM)*, (pp. 1-9), 2016.

[21] O. Goga, H. Lei, S. Parthasarathi, G. Friedland, R. Sommer and R. Teixeira, "Exploiting innocuous activity for correlating users across sites", *WWW'13*, (pp. 447-458), 2013.

[22] R. Zafarani and H. Liu. "Connecting users across social media sites: A behavioral-modeling approach", In *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*, (pp. 41-49), 2013.

[23] Y. Zhang, J. Tang, Z. Yang, J. Pei, and S. Yu, "COSNET: Connecting Heterogeneous Social Networks with Local and Global Consistency", *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, (pp. 1485-1494), 2015.

[24] M. Li, N. Cao, S. Yu, and W. Lou, "FindU: Privacy-Preserving Personal Profile Matching in Mobile Social Network", In *Proceedings of IEEE INFOCOM*, (pp. 2435-2443), 2011.

[25] M. Li, S. Yu, N. Cao, and W. Lou, "Privacy-Preserving Distributed Profile Matching in Proximity-based Mobile Social Networks", IEEE *TWC*, vol.12, no.5, 2013.

[26] R. Zhang, Y. Zhang, J. Sun, G. Yan, "Fine-grained private matching for proximity-based mobile social networking", In *Proceedings of IEEE INFOCOM*, (pp. 1969-1977), 2012.

[27] R. Zhang, Y. Zhang, J. Sun, G. Yan, "Privacy-Preserving Profile Matching for Proximity-Based Mobile Social Networking", IEEE *JSAC*, vol. 31, no. 9, pp. 656-668, 2013.

[28] M. Rosvall and C. T. Bergstrom. "Maps of random walks on complex networks reveal community structure", *PNAS'08*, 105(4):1118-1123, 2008.

[29] A. E. Monge, and E. Charles, "The Field Matching Problem: Algorithms and Applications", *Proceedings of the ACM SIGKDD international conference on Knowledge discovery and data mining*, (pp. 267-270), 1996.

[30] W. Cohen, P. Ravikumar and S. Fienberg, "A comparison of string metrics for matching names and records." *Kdd workshop on data cleaning and object consolidation*, Vol. 3, 73–78, 2003.

[31] D. Perito, C. Castelluccia, M. A. Kaafar, and P. Manils. "How unique and traceable are usernames?" In *PEST'11*, (pp. 1-17), 2011.

[32] L. Yartseva and M. Grossglauser. "On the performance of percolation graph matching", In *Proceedings of the first ACM conference on Online social networks*, (pp. 119-130), 2013.