International Journal of Research
Available at https://edupediapublications.org/journals

p-ISSN: 2348-6848
e-ISSN: 2348-795X
Volume 06 Issue 13
December 2019

Privacy Leakage Via De-Anonymization And Aggregation In Heterogeneous Social Networks

Mrs Reshma Mahjabeen M. SC (I S), Lecturer, King Khalid University

Abstract: Although the approach represents a personal commitment, guidance, and recommendation, user profile accumulation of several social networks would result in significant privacy breaches. In this proposal, we propose a Novel Heterogeneous De-anonymization Scheme (NHDS). The NHDS began using a graphical design of the network to dramatically reduce the size of the set of candidates and then used the information to identify the profile that the user mapping and the degree of belief were high. Evaluation of performance data in real social networks shows that NHDS significantly exceeds the original schedule. Finally, we conduct a scientific study of the loss of integrity of results raised several networks based on four sets of social network data. Our findings show that 39.9% of the information is captured in anonymity and the anonymization ratio is 84%. The second policy leaks the population and the interests of users also under consideration, indicating the potential for leaks is recognized privately..

Index Terms: Data privacy, Social networks security, De-anonymization, Heterogeneous social networks

1. INTRODUCTION

Social networks (online social networks, mobile social networks, vehicular social networks, etc.), or social media, have been extremely popular in current days. The latest statistics show that the number of active traditional social media users has exceeded 2.7 billion [1]. 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 [2]. 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 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. 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 [3]. 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]. This issue can not only exist in traditional social networks but also exist in new emerging social networks, like vehicular social networks. For example, Twittermobile car is able to send and receive Twitter messages, which contain the information including drivers' status, vehicle profiles, and real-time traffic notifications; Road Speak is a voice chatting system used by daily driving commuters or a group of people who are on a commuter bus or train [5]. These vehicular social-based applications exploit traditional online social networking services, like Facebook and Twitter, and thus are also under threat of de-anonymization attack. In this study, we take an initial step towards investigating the following two questions: i) How can we design a practical and effective cross-network aggregation scheme for heterogeneous social networks? The proposed cross-network aggregation scheme is expected to link the target user's various accounts on different social media platforms and collect the user's profile in different aspects. ii) To what degree the cross-domain aggregations can reveal the different attributes of a user (e.g. interest, demographics). One of the fundamental challenges of bridging the different social identities of the users on different social media is that the users tend to use varying usernames (screen names) 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 'deanonymization'. 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 assume the social networks are aligned.

In this study, to answer the above questions, we first present a Novel Heterogeneous De-anonymization Scheme for heterogeneous social networks, which is coined as NHDS. Different from any previous works which either focus on profile based or structure based approach, NHDS aims to integrate the merits of two kinds of approaches. The motivation is that a real-world attacker is able to leverage as much information as she can to help deanonymize in practice. Since both user profiles and network graph topology can be collected through web crawlers, platforms' APIs, or public datasets, a novel approach that leverages the merits of these two strategies is expected to achieve a higher performance. 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. To further investigate the privacy leakage caused by the cross network aggregation, we apply the proposed NHDS algorithm to a large dataset involving four real-world heterogeneous online social networks, i.e., Livejournal, Flickr, Last.fm, and MySpace. We perform the de-anonymization algorithm and measure the privacy leakage arising from cross-network aggregation. The results are quite surprising in that, with the proposed de-anonymization algorithm, cross-network aggregations can reveal 39.9% uncovered attributes of users (e.g. interest, demographics).

2. Literature Survey

2.1 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. 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. 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 [9].Nilizadeh et al. extended Narayanan and Shmatikov's attack by proposing a community-enhanced deanonymizing scheme of social networks. Then, Lai 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. Some papers modeled mobility traces as graphs and presented different attacks for de-anonymizing using online social networks as side channel. 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 patterns 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 deanonymize users. Besides, Ji et al. conducted the comprehensive quantification on the de-anonymizability of 24 real-world social networks with seed information in general scenarios. Later, in, a uniform and open-source secure graph data sharing/publishing system was proposed. Li et al. proposed a graph-based framework for privacy preserving data publish, which is a systematic abstraction of existing anonymity approaches and privacy criteria. Qian et al. leveraged background knowledge graph to improve the de-anonymization performance. But this work mainly focuses on de-anonymizing a graph anonymized from original graph and inferring some private attributes. Fu et al. proposed a graph node similarity measurement in consideration with both graph structure and descriptive information, and a deanonymization algorithm based on the measurement. Zhang et al. targeted Twitter users in a metropolitan area by exploiting the strong geographic locality within communications on Twitter. In our work, we try to de-anonymize heterogeneous social networks by considering both semantic information and structure information, and evaluate the privacy leakage after de-anonymization.

2.2 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, Stumble Upon and Flickr. Olga et al. extracted features and developed supervised machine learning models which can perform entity matching between two profiles for a user by similar name and deanonymizing a user's identity. 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. Vosecky et al. identified users between Facebook and StudiVZ by exploiting various profile attributes. Zafarani et al. 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. Wondracek et al. introduced a technique that narrows down user identity by examining social-network group membership stolen from browsing history. Zhang et al. 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 and, the first privacy-preserving personal profile matching scheme for mobile social networks was proposed by Li et al. In this scheme, an initiating user can be identified from a group of users the one whose profile best matches with his/her, with limited risk of privacy exposure. Later, two novel fine-grained private profile matching protocols were designed in. Different from these works, our proposed approach uses social structure to narrow down the candidate sets in order to achieve higher accuracy.

3. OVERVIEW OF THESYSTEM

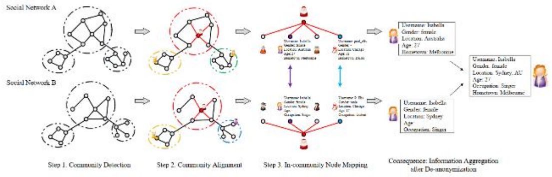


Fig 3.1System Overview

Proposed Work:

As we know, the work that is proposed is the first quantitative study to assess the impact of anonymization and aggregate network through the loss of integrity in real data. From a privacy protection point of view, our study also said the public's potential risk regarding anonymity was not registered and the collection of information and research needs was to retain personal advice. The main contribution of this document can be summarized as follows:

- We must consolidate many different systems to identify new users through multiple social networks. Collectively, the proposed program uses publicly available information about the network structure and user profile, which is expected to significantly increase accuracy.

- We perform a comprehensive set of experiments on many different data in social networks to show that the proposed program is the human species. The comparative results show that HDS achieve the accuracy of the information and record that key compared to the original value.

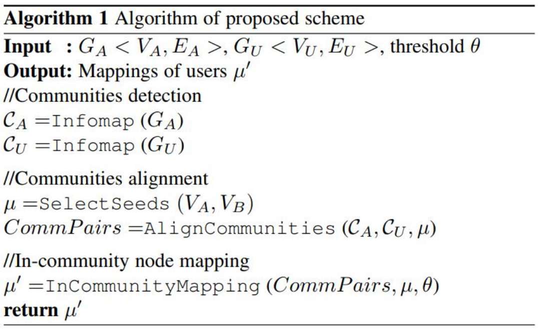
- In order to understand the impact of dance attack properties or field concentration, we investigate and measure the transmission of information through a network of general anonymous social networking sites. The result shows that 39.9% of the information is open and that the anonymous relationship (described in section 6.3) is 84%, which raised the issue of privacy seriously.

4.NOVEL HETEROGENEOUS DE-ANONYMIZATION SCHEME

In this section, we introduce our proposed Novel Heterogeneous De-anonymization Scheme (NHDS).

4.1 Scheme Overview

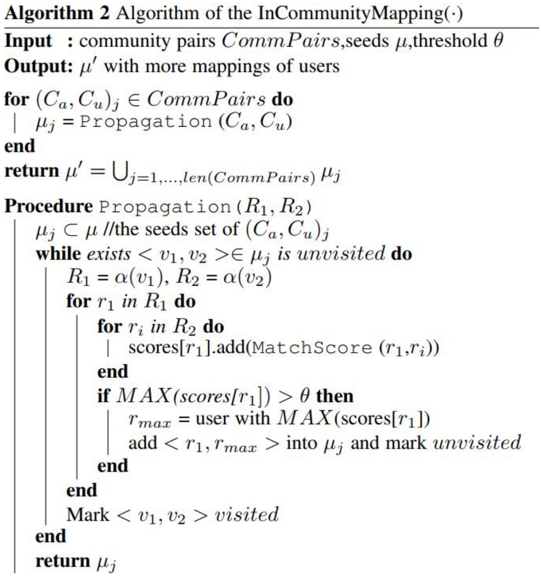
It illustrates our proposed scheme which has three main steps: (1) Communities Detection: communities in both networks are detected according to graph structure, (2) Communities Alignment: seeds are automatically identified based on profiles, and communities that contain the same pairs of seeds are aligned, (3) In-community node mapping: in each pair of aligned communities, nodes with high similarity score, which is computed by profile similarity, is accepted as a mapping, and mapping process is propagated to the neighbors. Algorithm. 1 presents the whole procedure, and the details and time complexity are introduced in the following sub-sections.



```
Algorithm 1 Algorithm of proposed scheme
Input  : $G_A < V_A, E_A >, G_U < V_U, E_U >$, threshold $\theta$
Output: Mappings of users $\mu'$
//Communities detection
$\mathcal{C}_A$ =Infomap ($G_A$)
$\mathcal{C}_U$ =Infomap ($G_U$)

//Communities alignment
$\mu$ =SelectSeeds ($V_A, V_B$)
$CommPairs$ =AlignCommunities ($\mathcal{C}_A, \mathcal{C}_U, \mu$)

//In-community node mapping
$\mu'$ =InCommunityMapping ($CommPairs, \mu, \theta$)
return $\mu'$
```



```
Algorithm 2 Algorithm of the InCommunityMapping($\cdot$)
Input  : community pairs $CommPairs$,seeds $\mu$,threshold $\theta$
Output: $\mu'$ with more mappings of users

for $(C_a, C_u)_j \in CommPairs$ do
    $\mu_j$ = Propagation ($C_a, C_u$)
end
return $\mu' = \bigcup_{j=1,...,len(CommPairs)} \mu_j$

Procedure Propagation($R_1, R_2$)
    $\mu_j \subset \mu$ //the seeds set of $(C_a, C_u)_j$
    while exists $< v_1, v_2 > \in \mu_j$ is unvisited do
        $R_1 = \alpha(v_1)$, $R_2 = \alpha(v_2)$
        for $r_1$ in $R_1$ do
            for $r_i$ in $R_2$ do
                scores[$r_1$].add(MatchScore ($r_1$,$r_i$))
            end
            if $MAX(scores[r_1]) > \theta$ then
                $r_{max}$ = user with $MAX$(scores[$r_1$])
                add $< r_1, r_{max} >$ into $\mu_j$ and mark $unvisited$
            end
        end
        Mark $< v_1, v_2 >$ $visited$
    end
    return $\mu_j$
```

5. EVALUATIONS OF PROPOSED SCHEME

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

5.1 Datasets

The datasets of four real-world heterogeneous online social networks, i.e., Live journal, Flickr, Last.fm, and MySpace, are obtained from. The datasets include node information, edge information, and profile information of a subset of users of these social networks.

- 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.

- Flickr is an image hosting online community for sharing, storing, and organizing photos. The dataset consists of

215,495 users and 9,114,557 friend relationship.

• 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.

• 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. These social networks not only provide different services and have different utility, but also have different graph properties. For example, Flickr has an average degree of 85.59 while the average degree of Livejournal is only 15.19. The heterogeneous structure increases the difficulty of de-anonymization.

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

TABLE 1: Statistics of social networks

Network	Nodes	Edges	Av. Degree
Livejournal	3,017,286	19,360,690	12.83
Flickr	215,495	9,114,557	85.59
Last.fm	136,420	1,685,524	24.71
Myspace	854,498	6,489,736	15.19

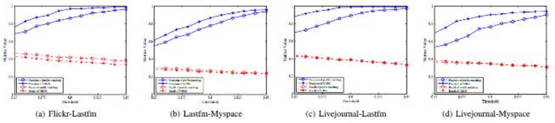

(a) Flickr-Lastfm (b) Lastfm-Myspace (c) Livejournal-Lastfm (d) Livejournal-Myspace

Fig. 1: Performance comparisons between profile-based matching and proposed NHDS

6. CONCLUSION AND FUTURE ENHANCEMENT

In this paper, we propose a practical Novel Heterogeneous Deanonymization Scheme (NHDS) for de-anonymizing real-world heterogeneous social networks, and evaluate and quantify the following privacy leakage. NHDS is a de-anonymizing scheme that exploits the network graph structure to significantly reduce the size of candidate set, and use user profile information to identify users with a high confidence. The performance evaluations of NHDS based on a dataset of four real-world social networks show that it achieves a high precision with a slight sacrifice of recall. We further quantify privacy leakage through de-anonymization. Evaluations show that notable portions of user information is disclosed. Privacy preserving in social networks is still an open challenge.

7. REFERENCES

[1] D. Chaffey, "Global social media research summary 2016", 2016. http://www.smartinsights.com/social-media-marketing/social-mediastrategy/new-global-social-media-research/

[2] 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/.

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

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

[5] A. M. Vegni, V. Loscri, "A survey on vehicular social networks," IEEE Communications Surveys & Tutorials, 17(4), 2397-2419, 2015.

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

[7] Z. Zhang, Q. Gu, T. Yue, and S. Su, "Identifying the same person across two similar social networks in a unified way: Globally and locally" In Information Sciences, 394, 53-67, 2017

[8] P. Pedarsani, D. R. Figueiredo, and M. Grossglauser. "A bayesian method for matching two similar graphs without seeds". In 51st Annual Allerton Conference on Communication, Control, and Computing (Allerton), (pp. 1598-1607), 2013.

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