A De-anonymization Attack for Social Network Graph Based on Structural and Node Feature Similarity

Qin Chen, Zhaoyong Wang, Min Zhang, Honglin Zhu and Shijie Feng

ABSTRACT

With the wide usage of Internet, social network has become an important carrier of information publishing and transmission in contemporary society. However, the anonymized information can still be de-anonymized because of the high frequency data-sharing in social network, which causes the leakage of users’ privacy information. This paper proposes a novel de-anonymization method based on the structure of the social network graph of users and the characteristics of user nodes to disclose users’ private information. It utilizes the similarity difference between the structural features of the social network graph and the characteristics of the node attributes to realize the mapping between nodes, so as to successfully de-anonymize the users of anonymous datasets. Experiments on two real datasets show that our anonymization method outperforms in accuracy and execution time.

KEYWORDS

Social Network; Privacy; Similarity; De-anonymization.

INTRODUCTION

Nowadays, the social network has played an indispensable role in people's daily
life and work. However, the privacy and security issues are attracting more and more attention from researchers and data providers in social network data publishing [1-2]. Various privacy attacks and protection techniques have been proposed, including privacy protection based on k-anonymization technology [3-4] and map-based de-anonymization (DA), e.g., as [5].

The de-anonymization attack has been studied for a long time. The de-anonymization research is more based on the social network users' interest, location and relations and so on. In [6], Gam S. propose a mobility model based on a mobile Markov chain (MMC) to achieve this attack. Wei et al. [7] proposed a location-based point-of-interest identification method that describes the characteristics of the user's individual points of interest by extracting the characteristics of the user community. In addition, a variety of structure-based approaches have been proposed to de-anonymize social network datasets [8-9]. In [10], Narayanan propose a completely network-based topology method to perform de-anonymization attack. In paper [9], Ji et al. study the quantification, practice and implications of structured data de-anonymization. Using a network graph of friends to map mathematical models across multiple social networks to identify all accounts that belong to the same user, e.g., as [11].

Unfortunately, most of the previous works have some limitations: 1) Most of the previous attacks only focus on de-anonymity; that is, the mapping of anonymous user datasets and auxiliary users datasets. 2) An attacker typically does not consider the attribute characteristics of the graph node, and rarely infer the sensitive attributes according to correlations among attributes.

In this paper, we construct a graph model for de-anonymity attacks and use the model to describe the inference process of some sensitive attributes. Our goal is to achieve the mapping between nodes by similarity difference. This paper presents a de-anonymization method based on its structural and nodes feature in social network. It mainly analyzes the structural features and node characteristics in graphs, that is, a method based on the similarity between nodes. In addition, we can make some inferences about user-sensitive attribute or behavior-prone speculation on the updated prior attack graph, which helps us to mine user privacy information.

ATTACK MODEL

In this paper, we take both structural feature and node feature into consideration. We model the anonymized structural data by a graph $(G^a = (V^a, E^a))$, where $V^a = \{i | i$ is an anonymized user$\}$ is the anonymized set, and $E^a = \{e(i,j) |$ is a link or relationship between $i \in V^a$ and $j \in V^a \}$ is a link set. Similarly, the auxiliary data is used to build an attacker's prior attack graph. The auxiliary data set can be expressed as $G^u = \{V^u, E^u\}$, where $V^u = \{i | i$ is a known user$\}$ and $E^u = \{e(i,j) |$ is a link between nodes, $i \in V^u$ and $j \in V^u \}$ is a link set. Note that the algorithm designed in this paper can be directly derived to apply to the directed graph scene.
Given the graph of two datasets ($G^a$ and $G^u$), de-anonymization attack schemes can be formally defined as the implementation of the mapping $\sigma: V^a \rightarrow V^u$. Based on the structural features of graph, we achieve the mapping between users as much as possible, and then identify the anonymous users.

We define some structural features for $i \in V^a$ or $V^u$ in this section, the specific definition as follows:

Degree: for $i \in V^a$ or $V^u$, its degree feature denoted as $d(i)$, which is its degree in $G^a$ (resp., $G^u$), i.e., $d(i) = |N_i^a|$ (resp., $|N_i^u|$).

Neighborhood: for $i \in V^a$ (resp., $G^u$), its neighborhood feature $n(i)$ is a $\beta$ - dimensional vector $(d_1^i, d_2^i, \cdots, d_\beta^i)$, where $\beta$ is a user-defined parameter (a non-negative integer) and $d_k^i (1 \leq k \leq \beta)$ is the kth largest degree in its neighborhood.

Top-K distance: for $i \in V^a$ or $V^u$, we define the Top-K distance feature $l_K^i(i)$ is a $K$-dimensional vector $(h_1^i, h_2^i, \cdots, h_\lambda^i, \cdots, h_K^i)$, where $h_k^i$ is the distance of a shortest path from $i$ to the user with the kth largest degree $(1 \leq \lambda \leq K)$. Note that it is possible $h_k^i = \infty$ if there is no edge relationship in the structural graph.

Sampling closeness centrality: for $i \in V^a$ or $V^u$, its sample closeness centrality feature $C(i)$ to describe its global topological feature, but do not result in much computational overhead. In fact, we randomly sample a subset $S^a \subseteq V^a$ and then we define $C(i) = \sum_{j \in S^a} \frac{1}{h(i,j)}$ (resp., $S^u \subseteq V^u$, $C(i) = \sum_{j \in S^u} \frac{1}{h(i,j)}$), where $h(i,j)$ is the distance from $i$ to $j$.

Taking into account the global structural features, we also define the Top-K reference distance and sampling closeness centrality, and analyze the structural characteristics of the graph from different perspectives, and then effectively distinguish the nodes, which helps to improve the accuracy of de-anonymization. Therefore, we first define a structural similarity as Formula (1):

$$S_s = c_1 \cdot s(d, c(i), d, c(j)) + c_2 \cdot s(n(i), n(j)) + c_3 \cdot s(l_K^i(i), l_K^j(j))$$  \hspace{1cm} (1)$$

Where $i \in V^a$, $j \in V^u$ and $c_1 + c_2 + c_3 = 1$, ($c_1, c_2, c_3 \in [0, 1]$ are constant value representing the weights), and $s(.)$ is the Cosine similarity between two vectors, and $d, c(i) = (d(i), c(i))$.

In addition to considering structural similarity, we also take into account the similarity of attributes between nodes. In terms of graph, each user node $i \in V^a$ has some links connection to attribute nodes. Given node $i$ and its occurrence attribute link with confidence score $c_s$, node $j \in V^u$, that is, similar to the above definition. Thus, the attribute similarity between nodes $i$ and $j$ is defined as the probability form $S_a(i, j)(2)$:

$$S_a(i, j) = \Pi_p c_s \cdot \Pi_{1-p} (1 - c_s)$$  \hspace{1cm} (2)$$

$$S(i, j) = W_s \cdot S_s(i, j) + (1 - W_s) \cdot S_a(i, j)$$  \hspace{1cm} (3)$$
In order to improve the accuracy of mapping between nodes, we should consider the structural features and node characteristics in the process of graph node mapping. Finally, we assign weights to $S_s(i, j), S_a(i, j)$, so the similarity of nodes $i$ and $j$ is $S(i, j)$ in formula (3).

Algorithm: Advanced De-anonymization

Require: Anonymized graph $G^a$, prior attack graph $G^u$, parameter $k, \theta$

Remove the isolated node $j \in G^a$.

1: Pick an initial node $i_0 \in V^u$ with high degree
2: Perform BFS nodes in $G^u$ from the starting node $i_0$
3: for each $i \in V^u$, following the BFS order of traversing $G^u$ do
4: for each $j \in V^a$ do
5: Calculate similarity $S(i, j)$, then get the candidate mapping
Set $c(i) = \text{GetTopSimilarity}(i, G^a, k)$
6: Employ the consistent rule to guarantee no mapping conflict, for $\sigma \in \prod_{i \in V^u}(i \times c(i))$, find the possible optimum mapping scheme
7: for each $(i, j) \in \sigma$, if the node similarity $S(i, j) \geq \theta$ then
8: accept the mapping $\sigma(i, j)$
9: if no mapping in $\sigma$ is accepted, break;

EXPERIMENT

We not just conduct de-anonymization attack experiments on two real world social network datasets, but we perform some privacy inference about user’s private attribute. And then we evaluate the performance of our de-anonymization method.

Datasets

We utilize our de-anonymization methods on two real datasets (Twitter and Google+) as shown in Table I. Firstly, as to twitter, we obtain their profiles and network data from 1000 ego-networks, consisting of 81306 nodes and 1768149 edges and users with incomplete or invalid error attributes should be removed. For Google+, we gain users’ data from 133 ego-networks, consisting of 107614 users. In addition, duplicate and isolated user profiles should be removed when we preprocess them.

Experimental Evaluation

In our experiments, we use its accuracy (the correct recognition of the user) and the run time as a measure of utility and complexity of the standard indicators. Firstly, we focus on the dataset Google+. The default parameter is set to: $k = 10, \theta = 0.7$ or $0.8, W_s = 0.7$. All the experiments are implemented on a PC with 64 bit Ubuntu 14.04 LTS operating system, Intel Xeon CPU (2.4GHz × 8 Threads),
and our experiment is repeated many times (since $G^a$ and $G^u$ are randomly generated) and the results are the average values.

### Table I. Relative Statistics of Two Datasets.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Type</th>
<th>Nodes</th>
<th>Edges</th>
</tr>
</thead>
<tbody>
<tr>
<td>Google+</td>
<td>directed</td>
<td>107614</td>
<td>13673453</td>
</tr>
<tr>
<td>Twitter</td>
<td>directed</td>
<td>81306</td>
<td>1768149</td>
</tr>
</tbody>
</table>

**Figure 2.** The impact of parameters $k$ & $\theta$ on accuracy and runtime.

**Figure 3.** Impact of parameter $W_s$ on accuracy and runtime.

Figure 2 effectively demonstrates how the setting of parameters $k$ and $\theta$ affect the accuracy and time complexity of the de-anonymization. As shown in Figure 2 (a), the de-anonymization accuracy of the anonymous dataset increases with an increase in the value of $k (k \leq 10)$ (herein, we match a user in $G^u$, we select $k$ themost similar user nodes from $G^a$). But then, the de-anonymization accuracy hastiny fluctuations when $k > 10$. And when the $k$ value increases, the running time increases constantly. Therefore, we set $k = 10$ by default. As depicted in Figure 2
(b), the change of the similarity threshold $\theta$ has some different impact on the accuracy and running time. The setting of these parameters is very relevant to the balance of de-anonymization accuracy and run time.

In addition, some parameter settings also have a significant effect. As shown in Figure 3, when $0.5 \leq W_s \leq 0.8$, the de-anonymization method works better (note that $W_s, (1-W_s)$ are the weights assigned to the structural similarity and the attribute similarity). Thus, it also shows that the above two feature values play an important role in measuring the similarity between nodes.

Similar experiment on Titter, its de-anonymization accuracy is slightly lower. Why do Titter dataset experiments have similar results? This is because we use the Breadth First Search (BFC) to traverse the nodes and perform a k-anonymity processing on the node attributes. As we have considered, if there are more graph-based structural features (such as Sampling closeness centrality, Top-K distance, 1-hop neighborhood, etc.) into our method, which will improve our de-anonymization accuracy rate.

CONCLUSION & FUTURE WORK

In this paper, we propose a novel de-anonymization method based on the structure of the social network graph and the characteristics of nodes to disclose users’ private information. It comprehensively considers the difference between the structural features and the characteristics of the node attributes. The similarity matching between the user nodes achieves the mapping, and effectively infers some of the user's sensitive information. The experimental results show that our anonymization method outperforms in accuracy and execution time. Future work will focus on the following: 1). the fact that most of the existing anonymous technologies are susceptible to DA attacks based on graph structure, it is recommended to design effective solutions for this attack; 2). due to the importance of data publishing security, how to develop a secure platform for data publishing in the future, which will not be affected by semantic and structure-based DA attacks.

ACKNOWLEDGMENTS

This work is supported by Zhejiang Provincial Technical Plan Project (No. 2017C01064, 2017C01055, 2018C01088).

REFERENCES