Research Article

Structural Attack to Anonymous Graph of Social Networks

Tieying Zhu, Shanshan Wang, Xiangtao Li, Zhiguo Zhou, and Riming Zhang

School of Computer Science and Information Technology, Northeast Normal University, Changchun 130117, China

Correspondence should be addressed to Riming Zhang; zhangrm280@nenu.edu.cn

Received 19 June 2013; Revised 7 October 2013; Accepted 18 October 2013

Academic Editor: Siddhivinayak Kulkarni

Copyright © 2013 Tieying Zhu et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

With the rapid development of social networks and its applications, the demand of publishing and sharing social network data for the purpose of commercial or research is increasing. However, the disclosure risks of sensitive information of social network users are also arising. The paper proposes an effective structural attack to deanonymize social graph data. The attack uses the cumulative degree of \( n \)-hop neighbors of a node as the regional feature and combines it with the simulated annealing-based graph matching method to explore the nodes reidentification in anonymous social graphs. The simulation results on two social network datasets show that the attack is feasible in the nodes reidentification in anonymous graphs including the simply anonymous graph, randomized graph and \( k \)-isomorphism graph.

1. Introduction

A social network is a social relation structure which is made up of a set of social entities and their social ties or interactions. The research of social network analysis can be traced back to the contributions of Moreno who discusses the dynamics of social interactions within groups of people [1]. Nowadays, with the emerging of online social networks and services, the social relationship in reality has been extended to the virtual network world. Billions of users use online social website making friends, sharing pictures, micrologging and so on. The social structure hidden in the social network data are valuable for social analysis for the purpose of commerce or academy. For example, the user behavior and interests derived from social data are important for all the commercial recommendation systems [2, 3]. At the same time, more and more attentions are being paid to the privacy preservation problems in the process of using social networks and sharing social data, since data publishing and exchange increase the risk of disclosure and leakage of personal information of social network users [4, 5].

Social networks are usually modeled as graphs, in which the vertices represent social entities and the edges represent the social links or social ties. The properties of entities, such as age, gender, and SIN, can be represented as the attributes of vertices, and the properties of links between entities, such as the tightness of social ties, can be shown as the edge label or weight. Therefore, the natural and simple way to prevent the disclosure of personal information of social users is to remove the user portfolios, such as names and ISN, or replace them with random identifications. But the simple method cannot prevent the disclosure of personal sensitive information. The earliest privacy event causing public attention is the publishing of email data set of Enron Corpus [6]. Although the original purpose is for legal investigation, the regularity of email communications among employees within the company, even the organization structure of Enron Corpus, can be inferred from the email data. Other personal information disclosure events include the AOL Company publishing anonymized user search data for the research in search engine [7], and Netflix Company publishing user movie scoring data for improving the movie recommendation systems [8]. All of the intended purposes of these data publishing issues are not to leak users’ information, but it results in the privacy risks.

On the other hand, many privacy-preserving methods have been put forward and examined including \( k \)-anonymity based privacy preservation via edge modification, probabilistic privacy preservation via edge randomization and privacy preservation via clustering and generalization (see the recent review papers [9–11]). Besides these methods, the differential privacy method, which depends on specific
privacy guarantees and aims to make users in released data computationally indistinguishable from most of the other users in that data set, are paid more and more attention recently [12, 13].

In the paper, we present a structural attack method to de-anonymize social graph data, called $n$-hop neighbor Feature for Node Reidentification ($n$-hop neighFNR). The method relies only on the network structure. It uses the cumulative degree of $n$-hop neighbors as the regional feature and combines with the simulated annealing-based graph matching method. With the aid of auxiliary graph, it can reidentify the nodes in anonymized social graphs. The simulations on two data sets including Karate clubs [14] and email networks of URV [15] show it is feasible on de-anonymizing social graphs including the simple anonymous graph, randomized anonymous graph and $k$-isomorphism graph.

The rest of the paper is organized as follows. Section 2 presents the related work, and Section 3 describes the definition of $n$-hop neighbor feature and the node reidentification algorithm, followed by the experiments results on data sets in Section 4. Finally, we conclude the paper in Section 5.

2. Related Work

In the graph data of social networks, nodes usually correspond to the users in social networks and edges correspond to the relationship between users. The privacy attack to graph data of social networks aims to obtain the sensitive information including identity, friendship and other personal information that is hidden in social networks.

Backstrom et al. [16] firstly proposes the active and passive attack to simple anonymous social graphs. These two methods try to identify the target in the released social graph. The difference between them is whether the attackers change the graph data before data publishing. In active attacks, the adversary can create a certain number of Sybil nodes and edges linked to the target and embed these node and edges into the graph before data publish, then find these “Sybil” nodes together with the targeted users in the anonymized network. In passive attacks, attackers try to discover a target using their knowledge of local structure of the network around the target.

Different privacy attacks depend on different background knowledge. Zhou et al. [10] generalize some possible background knowledge that can be used in the privacy attacks. The background information includes degree, attribute of nodes, special links with the target node, neighbors, embedded sub-graphs, and other properties of graphs such as betweenness, closeness centrality and etc. Some literature [17–22] discusses different background knowledge and the corresponding privacy protection methods. Literature [17] proposes the known degree attack and corresponding $k$-degree anonymous graph for solving the problem. Literature [18] presents a degree trace attack which traces the change of certain node degree in the evolution graphs to reidentify the target node. Other privacy attacks are based on the structure of 1-hop neighbor [19] or neighbor subgraph and the corresponding privacy preserving methods are usually based on $k$-anonymity methods in structure, such as $k$-automorphism [20], $k$-isomorphism [21] and $k$-symmetry model [22]. Compared with these $k$-anonymity methods, edge Randomization is a generalized privacy preservation methods which is not specific to the privacy attacks.

The de-anonymizing social network based on the auxiliary graph is a feasible attack which can recognize nodes from the large scale social networks. The auxiliary graph which can be obtained by crawling, is used to match the anonymized graph and reidentify the targets on the viewpoint of graph structure. Literature [23] proposes this method and use it to reidentify a third of the users who have accounts on both Twitter and Flickr with small error rate. In such attack, there are usually two phases: the recognition of certain amount of seed nodes in the anonymous graph, and then the propagation process to match the rest of nodes in the auxiliary graph with the targets in the anonymous graph on the basis of the known seeds. Rattigan [24] employs the crawling data of Yahoo! Music data as the auxiliary graph and attends to recognize the artist in the data set of KDD Cup 2011. Although there is no ground-truth graph to show the accuracy of this work, it still shows the feasibility of such attack. More recently, Narayanan et al. [25] use the crawling Flickr graph to deanonymize much of the competition test set of a machine learning contest in Kaggle challenge. They use the neighbor similarity between node pairs as a structural feature and combine it with simulated annealing-based graph matching method to reidentify a small number of nodes with largest degree. These recognized nodes can be used as the seed nodes in the first phase of de-anonymizing attack.

In privacy attacks mentioned above, the degree, neighbor structure, neighbor similarity of certain node pair are all metrics used to match or recognize the target nodes in anonymized graphs. In graph mining, Henderson et al. [26] present the concept of recursive feature, which combines node-based local features with egonet-based neighborhood features to capture the regional information of an individual node in large graphs. It can be applied in de-anonymization tasks on evolution graphs or partially anonymized graphs. Influenced by these ideas, the paper proposes a structural attack, which combines the cumulative $n$-hop neighbors’ degree feature with the simulated annealing algorithm, to reidentify the nodes in anonymized graphs.

Many computational intelligent algorithms for optimization problems are proposed like in literature [27–31]. In this paper, we use simulated annealing method [32] to match the auxiliary graph with anonymized graph, although other intelligent algorithms can be used to replace simulated annealing method.

3. $n$-Hop Neighbor Feature for Node Reidentification

3.1. $n$-Hop Neighbor Feature. A social network can be modeled as a graph $G = (V, E)$, where $V$ is the set of nodes and $E$ is the set of connections. In this paper, undirected graph is used, although social networks can be directed graphs if the direction of connections is considered. In a graph structure,
the node degree is the basic local feature of a node. When considering the relation of a node and its 1-hop neighbors, the related metrics are computed in the range of egonet. For example, in literature [25], the cosine similarity between a pair of nodes is defined as: $|x ∩ y|/\sqrt{|x| \cdot |y|}$, while $x$ and $y$ are the neighbor sets of a node pair, respectively.

In this paper, we considered the cumulative degree feature of a node in the range of $n$-hop neighbors. For a node $i$, its cumulative degree is defined as the sum of $n$-hop neighbors’ degree of a node and denoted as follows:

$$\sum_{j_1 \in \text{neigh}(i)} \sum_{j_2 \in \text{neigh}(j_1)} \ldots \sum_{j_n \in \text{neigh}(j_{n-1})} \text{degree}(j_n).$$

(1)

The $n$-hop neighbor feature is a regional feature and captures the node's properties better than the node degree since for the nodes with same degree would have different value of cumulative value $n$-hop neighbor degree. It qualifies the connections of a node with other nodes and shows the importance of a node in the $n$-hop range and even the whole network.

In order to show the discrimination of the $n$-hop cumulative feature, we analyze two data sets: karate club data and email network of URV, in terms of the value of degree and $n$-hop cumulative feature; here, $n = 4$. In karate data set, there
there are 11 nodes with the same degree 2, while there are 5 nodes with same \( n \)-hop neighbor feature value of 6376 and another 2 nodes with the other \( n \)-hop neighbor feature of 5977. These figures show that most of the nodes with the same degree have different 4-hop neighbor feature value on the two data sets. Specially, Figure I(b) shows that most of the nodes with large degree value in email network data can be discriminated by 4-hop neighbor feature value. Although the discrimination becomes worse for some lower degree value, like 3 or 2, it is still better than the degree feature.

3.2. \( n \)-Hop Neighbor Feature for Node Reidentification. The \( n \)-hop neighbor feature can capture the regional information of a node in the graph. This paper combines \( n \)-hop neighbor feature with simulated annealing algorithm and proposes \( n \)-hop neighbor feature for node reidentification algorithm (\( n \)-hop neighFNR) to deanonymize social networks by the aid of auxiliary graph.

For two graphs \( G_A \) and \( G_T \), \( G_A \) is the auxiliary graph, in which the identities of nodes are already known, and \( G_T \) is the anonymous target graph. \( G_A \) and \( G_T \) can be thought as the induced graph from the same graph \( G \). Usually the auxiliary graph \( G_A \) can be obtained by the crawling. \( G_T \) is the anonymous social data for publishing. The process of privacy attack can be thought as a matching process between the nodes of \( G_A \) and \( G_T \). We use the original data sets as \( G_A \), and three kinds of anonymous graphs as \( G_T \), including the simple anonymous graph, which removes the identification of nodes, the randomize anonymous graph, which is obtained by randomly adding one edge followed by deleting another edge and repeating the process for \( k \) times, as shown in [33], and \( k \)-isomorphism graph, as shown in [21].

We combine the \( n \)-hop neighbor feature with the simulated annealing methods to match the two graphs, \( G_A \) and \( G_T \). The Euclidian distance between two sets of \( n \)-hop neighbor feature value of node pairs is defined to measure the quality of a candidate mapping, so that we can optimize the matching over all the possible mapping. The Euclidian distance is defined as:

\[
\sqrt{\sum_{i=1}^{n} (x_i - y_i)^2},
\]

in which \( x_i \) and \( y_i \) is the cumulative degree value of node \( i \)'s \( n \)-hop neighbors in \( G_A \) and \( G_T \), respectively.

Algorithm I shows the method in \( n \)-hop neighbor feature for node reidentification. In the algorithm, \( \Delta p \) is the change of Euclidian distance in different matching processes. \( T \) is the temperature, which will be cooled with a rate of \( \alpha; \alpha = 0.9 \). The initiating value of temperature depends on the nodes in the graph; \( T = 100 \ast n \). The ending of simulated annealing algorithm is determined by the threshold \( T_{min} \). \( c \) is constant, which is also dependent on the nodes, and \( c = 20 \ast n \).

4. Experiment Results

In order to show the feasibility and effectiveness of \( n \)-hop neighbor feature, we compare it with the neighbor similar feature used in literature [25], which is called neighSNR, to deanonymize the simply anonymous graphs, randomized graphs, and \( k \)-isomorphism graphs. In e-mail data set, we select 30 nodes with the largest degree.

**Figure 2:** The results on simple anonymous graph for karate data set and e-mail network data.

are 34 nodes and 156 edges, while there are 1133 nodes and 5451 edges in email data set. Figures 1(a) and 1(b) shows that \( n \)-hop neighbor features discriminates nodes much better than the feature of degree on both two data sets. In these figures, \( x \) presents the degree of nodes in decreasing order, and \( y \) presents the corresponding 4-hop neighbor feature value. In a graph, there may be some nodes with the same degree value, and the nodes with the same degree value may also have the same or different 4-hop neighbor features; therefore we use the number besides the circle to denote how many nodes have the same \( n \)-hop neighbors feature value. For example, in Figure I(a), in the 5 nodes with same degree 4, 3 nodes have different \( n \)-hop neighbor feature value and only 2 nodes have same \( n \)-hop neighbor feature value. Furthermore,
as the targets. The randomized graphs are obtained by adding or deleting edges randomly and repeatedly. The perturbation degree is defined as the percentage of the number of adding/deleting edges and we use 10%, 20%, and 50% as the perturbation degree, respectively. The $k$-isomorphism graph is obtained by adding or deleting edges to satisfy the definition of $k$-isomorphic: A graph $G$ is $k$-isomorphic if $G$ consists of $k$ disjoint subgraphs $g_1, \ldots, g_k$, where $g_i$ and $g_j$ are isomorphic for $i \neq j$. In the experiment, we use 2-isomorphic graph as the anonymous graph [21].

Figures 2(a) and 2(b) show the recognition results in simple anonymous graph. Our method $n$-hop neighFNR is much better than neighSNR. It can recognize 32 out of 34 nodes for karate data set in the best case and all the 30 largest degree nodes for email network data.
Input: $G_A$ and $G_T$: the auxiliary graph and target anonymous graph
Output: node mapping between $G_A$ and $G_T$
Initialize the original mapping and temperature $T$;
while (Temperature $T >$ Threshold $T_{min}$)
  Exchange two nodes position to make a new mapping;
  if (the Euclidian distance of the new mapping < the Euclidian distance of old mapping)
    Accept the changing of nodes position;
  else
    Accept it with a probability $e^{-\Delta p/c*T}$;
  endif
  Decrease the temperature $T$ with a rate;
endwhile

Algorithm 1: $n$-hop neighbor feature for node reidentification.

Figures 3(a) and 3(b) show the results of the number of reidentification node in randomized anonymous graph with different perturbation degree for karate data and email network data. Although with the increasing of perturbation degree, the number of reidentification nodes decrease, our algorithm neighFNR outperforms neighSNR in general. For karate data set, 12 node pairs are matched in the best case when our method is employed and perturbation degree is 10%. When perturbation degree increases to 50%, 8 node pairs are re-identified using our algorithm. For neighSNR method, 4 node pairs are recognized when perturbation degree is 50% in the best case. For email network data, when perturbation degree is 10%, 10 nodes pairs are matched. When perturbation degree increases to 50%, both the $n$-hop neighFNR and neighSNR method recognize 4 nodes pairs in the best case. In the average and worst cases, $n$-hop neighFNR also outperforms neighSNR on both of two data sets.

Figures 4(a) and 4(b) show the de-anonymizing results on $k$-isomorphism graphs, in which $k = 2$. In the graph of Karate data, 16 edges are added and 36 edges are deleted in order to generate 2-isomorphism anonymous graph. In the graph of Email data, 2428 edges are added and 2454 edges are deleted. Both neighSNR and $n$-hop neighFNR does not work well on 2-isomorphism graphs, since these two 2-isomorphism anonymous graphs are obtained by perturbation of about 50% edges of the corresponding original graphs, and the $k$-isomorphism method enforces $k$-security for protecting the nodes and links in anonymous graph [21].

5. Conclusion
The paper presents $n$-hop neighbor feature to capture node characteristics in a graph. It uses the sum of degree of $n$-hop neighbors of a node as a regional feature. When combining with simulated annealing algorithm, it can be used as a structural attack to de-anonymize social networks. The experiments on two data sets show it is very effective for de-anonymizing the simple anonymous graph and feasible
for the randomized graph. The research provides insights for the privacy-preserving problem of social networks and the design of privacy-preserving algorithms. The future work we should do is to evaluate the effectiveness of our algorithm on large-scale real networks.

Acknowledgments

This work was supported in part by the Special Fund for Fast Sharing of Science Paper in Net Era by CSTDF (FSSP 2012) and the Technical Development Plan of Jilin Province of China (No. 201101003).

References


