A Privacy Protection Method for Social Networks Based on Node Importance

CUI Haitao
School of Computer Science
Nanjing University of Posts and Telecommunications
Nanjing 210023, China
chtnjupt@163.com

LI Lingjuan
School of Computer Science
Nanjing University of Posts and Telecommunications
Nanjing 210023, China
Jiangsu Key Laboratory of Big Data Security &
Intelligent Processing
Nanjing 210023, China
lilj@njupt.edu.cn

***Abstract*—The social networks record a large amount of social activities information, which contain sensitive data and private information. It makes social networks more vulnerable to be attacked. In order to resist the attacks based on structural knowledge and better protect the privacy information, this paper designs a privacy protection method for social networks based on node importance (NI-PP). This method introduces the node importance calculated by K-shell into K-symmetric anonymity to realize the purpose of protecting important nodes and provides a new restoration algorithm that can completely restore the structures of the original networks. This restoration algorithm uses labels added on the original networks to recover those nodes that are deleted mistakenly during the restoration process, so as to improve the availability of the anonymous networks and reduce the loss of information in the restoration process. Analysis and experiments results show that NI-PP can protect the privacy information of social networks very well. (*Abstract*)**

Keywords—social network; privacy protection; K-shell; K-symmetric anonymity (key words)

I. INTRODUCTION

With the development of Internet, the personal privacy of users become more and more important. In social networks, every user is socializing, and the main purpose of socializing is to share information. In the process of information sharing, the information will be encrypted and hidden in an anonymous way. However, this approach is not suitable for social networks. If the information is hidden, the user cannot complete the socializing effectively [1]. And some of the information are sensitive, which carry privacy. Therefore, many people are motivated by interests to attack such data sets which contain private information and use the obtained private information to conduct illegal activities such as fraud on users.

In order to avoid the problem of privacy disclosure, many scholars have studied the privacy protection and proposed many methods such as K-anonymity [2], K-symmetric anonymity and differential privacy protection [3], etc. To improve the efficiency of privacy protection for social networks, this paper focuses on the attacks which are based on structural knowledge and the protection of important nodes. By improving K-symmetric anonymity and designing a new restoration algorithm, this paper proposes a new privacy protection method for social networks based on node importance and analyzes the availability and confidentiality of the method.

II. RELEVANT KNOWLEDGE

A. *Types of Attacks in Social Networks*

According to the strategies adopted by attackers, the attacks on social networks can be divided into three types: active attacks [4], passive attacks [5] and attacks based on background knowledge [6].

1) Active Attacks: Active attacks mean that before the anonymous network is released, the attacker has learned that the target node exists in the network. The target node is marked by adding nodes and edges before the anonymous network is released, and then after releasing the anonymous network, the attacker locates the target node according to the relevant markers and steals the private information.

2) Passive Attacks: Passive attacks mean that the attacker directly masquerades his or her own information in the network without adding nodes and edges before releasing the anonymous network. After the anonymous network is released, the attacker uses these masqueraded nodes to probe other nodes connected to them to implement the attacks and obtain private information.

3) Attacks Based on Background Knowledge: In fact, the above two attacks are difficult to implement. Therefore, more cases are that the attacker attacks the network after the network is released, but since there is no relevant markers for positioning, the attacker can only obtain the information of the target node through other ways to recognize and attack it. The information that an attacker can obtain is called background knowledge, so this type of attacks is called attacks based on background knowledge.

The typical attacks based on background knowledge are the attacks based on structural knowledge. It means that the attacker locates nodes through the topological structures of the networks to obtain private information. This type of attacks is very common in social networks.

B. *Anonymity Technology*

Anonymity is one of the important ways to protect privacy information.

1) Simple Symmetric Anonymity: Simple symmetric anonymity uses the method of replacing or hiding the identification information in the social networks without changing the topology of the original network to anonymize the networks, and then find nodes without symmetric nodes and replicate them symmetrically.

2) K-symmetric Model: The K-symmetric model uses the relevant theory of graph theory to replicate the nodes in the network, so that the number of records containing the same identifier in the copied data set is at least *k*, so the probability that the target is accurately identified is not higher than *1/k*.

3) K-symmetric Anonymity: In the attacks based on structural knowledge, it is easy for the attacker to distinguish the nodes without symmetrical nodes. Therefore, some researchers design K-symmetric anonymity [7] based on K-symmetric model. The basic steps of this method are anonymizing all nodes with simple symmetric anonymity, then anonymizing the nodes with K-symmetric anonymity. K-symmetric anonymity requires each node to have *k-1* nodes that are symmetrical with it, so that the probability of an attacker to identify the target node is not greater than *1/ k*, so as to resist the attacks.

III. Design of NI-PP

A. Idea of NI-PP

K-symmetric anonymity can be used to resist the attacks based on structural knowledge, but K-symmetric anonymous networks have many redundant nodes and edges, so the costs of anonymity are high. The researchers put forward some methods to reduce the addition of nodes and edges and decrease the costs of anonymity. However, many methods only consider the degree of nodes, and ignore the importance of nodes in other aspects. In fact, important nodes may dominate the network, so the privacy protection of important nodes should be strengthened. For those less important nodes, only the basic anonymity methods are used to anonymize them.

This paper designs a privacy protection method for social networks based on node importance (NI-PP) that considers the importance of nodes by introducing K-shell on the basis of K-symmetric anonymity firstly. Secondly, in order to ensure the availability of the anonymous networks, a restoration algorithm that can completely restore the structures of the original networks is designed.

B. K-shell method to measure the importance of nodes

In social networks, the important nodes often play a dominant role. There are many ways to measure the importance of nodes, such as degree centrality, median centrality, eigenvector centrality, k-shell [8], etc. But considering the efficiency of the algorithm and the attacker can easily obtain the degree of nodes, this paper uses the K-shell method to measure the importance of the nodes.

The purpose of the K-shell is to divide the nodes according to *ks* value. The specific dividing process is as follows: In the network without isolated nodes (the degree of nodes is 0), delete the nodes whose degree is 1 and the corresponding edges firstly, and then continue to delete those newly appeared nodes whose degree is 1 and the corresponding edges until there is no node with degree is 1 in the remaining nodes, so that all the deleted nodes form the first layer (1-shell), and their *ks*=1. Repeat the above operation on the remaining nodes to get the second layer of *ks*=2 (2-shell). And so on, until all the nodes in the network have the *ks* value.

C. Improvement of K-symmetric Anonymity

The specific steps of the improved K-symmetric anonymity are as follows.

Step 1. Obtain the adjacency matrix of the original network and add a label on each node which records the number of isomorphic nodes (nodes with the same row and column values).

Step 2. Carry out simple symmetric anonymity on the original network.

Step 3. Determine *k* value according to confidentiality requirements.

Step 4. Calculate *ks* values of all nodes by K-shell algorithm.

Step 5. Replicate the nodes with the maximum ks value in the adjacency matrix.

Step 6. Repeat ***Step 4*** and ***Step 5*** according to the *k* value until the condition is met.

Step 7. At the end of replications, transform the adjacency matrix into a graph.

D. Design of The New Restoration Algorithm

The original K-symmetric anonymity adds many new nodes and edges to the original social networks, so that there is a big difference of the topological structures between anonymous networks and original networks. In order to ensure the availability of anonymous networks and solve the problem that the ordinary restoration algorithm cannot completely restore the structures of the original networks, the specific steps of the new restoration algorithm are as follows.

Step 1. Analyze the adjacency matrix of the anonymous network, find out the same columns in the matrix and delete them in turn until there are no columns that are the same with them in the matrix.

Step 2. Find out the same rows in the matrix and delete them in turn until there are no rows that are the same with them in the matrix, and then get the initial restoration matrix.

Step 3. According to the labels added in the ***Step 1*** of the improved K-symmetric anonymity, review the initial restoration matrix. If the number of nodes is found to be inconsistent with the number in the labels, then carry out symmetrical replication until the number is the same as that in the labels. At this time, the obtained restoration matrix is the final result.

Step 4. Generate the original social network according to the final restoration matrix.

IV. Analysis of NI-PP Method

A. Confidentiality Analysis.

The improved K-symmetric anonymity makes each node has *k*-1 symmetrical nodes through replicating nodes symmetrically, it means that each set of the equivalent classes contains k nodes. The analysis of confidentiality in different situations is as follows.

1) Case 1: All n nodes in the original social network are isolated nodes, that is, no node has symmetrical nodes. At this time, the probability of cracking the anonymous network is shown in (1).

$$P_1 = \left(\frac{1}{k}\right)^n \tag{1}$$

2) Case 2: For an original social network with n nodes, it is assumed that each set of the equivalent classes contains at least t nodes. the probability of cracking the anonymous network is shown in (2).

$$P_2 = \left(\frac{1}{k}\right)^{\frac{n}{t}} \left(\frac{1}{k-1}\right)^{\frac{n}{t}} \left(\frac{1}{k-2}\right)^{\frac{n}{t}} \cdots \left(\frac{1}{k-(t-1)}\right)^{\frac{n}{t}} \tag{2}$$

For the different number of nodes in the equivalent sets, the confidentiality is also very different. For example, assuming that each result set contains two symmetrical nodes, the probability of cracking the anonymous network is shown in (3).

$$P_3 = \left(\frac{1}{k}\right)^{\frac{n}{2}} \left(\frac{1}{k-1}\right)^{\frac{n}{2}} \tag{3}$$

When the original network has only one equivalent set, that is, t=n, the probability of cracking the anonymous network is shown in (4).

$$P_4 = \left(\frac{1}{k}\right)\left(\frac{1}{k-1}\right)\left(\frac{1}{k-2}\right)\cdots\left(\frac{1}{k-(n-1)}\right) \tag{4}$$

In conclusion, the probability of cracking the anonymous network should be between P_1 and P_2. Taking n=20, k=8 as an example, its cracking probability is between $(1/8)^{20}$ and $(1/8)^{10}$×$(1/7)^{10}$. It can be seen that this algorithm has high confidentiality. Of course, these two cases are too extreme, in fact, the attackers may have some other information about the target node, so the actual probability of cracking should be slightly higher than the theoretical probability.

B. Availability Analysis.

The new restoration algorithm can restore the structures of the original social networks and reduce the loss of information in the restoration process.

V. Experiments And Results Analysis

A. Experiments Based on Simple Network

There are 8 nodes and 11 edges in Fig.1.

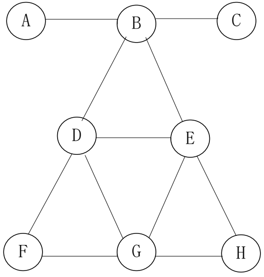

Fig. 1. The sample network

Carry out K- symmetric anonymity on the network until k=8, and the 8-symmetric anonymous network is shown in Fig.2.

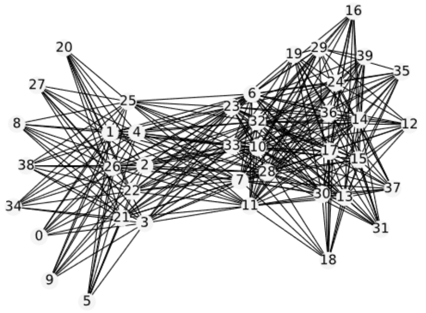

Fig. 2. The 8-symmetric anonymous network.

When k=8, each node has 7 symmetrical nodes, so the probability of identifying the target node is very low. But it has many redundant nodes and edges.

The improved K-symmetric anonymity is used to anonymize the simple symmetric anonymous network, and the result is shown in Fig.3.

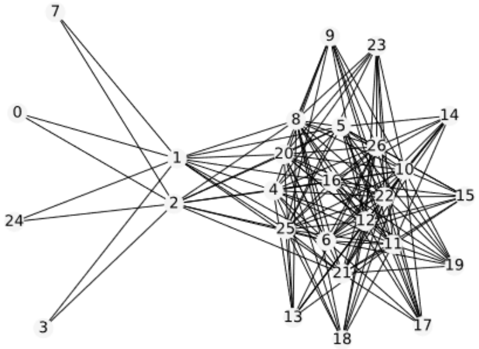

Fig. 3. The improved 8-symmetric anonymous network.

It can be found that compared with Fig.3 network, Fig.4 network has 13 fewer nodes and 152 fewer edges, which greatly reduces the costs of anonymity and ensures the protection for important nodes.

Considering the availability of the above anonymous network, the new restoration algorithm is used to recover it, and the result is shown in Fig.4.

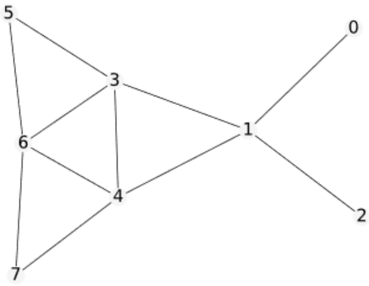

Fig. 4. The restoration result of improved 8-symmetric anonymous network.

It can be found that the result of the new restoration algorithm is the same as the structure of the sample network in Fig.1.

B. Experiments Based on Real Social Networks

This paper uses the improved K-symmetric anonymity algorithm to anonymize the three classic real social network data sets, which are Zachary Karate Club Network (with 34 nodes and 78 edges), Dolphin Network (with 62 nodes and 159 edges) and Football Club Network (with 115 nodes and 631 edges). And each network was anonymized 3 times continuously.

The results are shown in Table I and Table II.

TABLE I. COMPARISON OF NODES NUMBER OF REAL SOCIAL NETWORKS ANONYMOUS RESULTS

Algorithm	*K-symmetric Anonymity*	*Improved K-symmetric Anonymity*	*New restoration algorithm*
Node number of Karate	272	68	34
Node number of Dolphin	496	205	62
Node number of Football	920	888	115

TABLE II. COMPARISON OF EDGES NUMBER OF REAL SOCIAL NETWORKS ANONYMOUS RESULTS

Algorithm	*K-symmetric Anonymity*	*Improved K-symmetric Anonymity*	*New restoration algorithm*
Edge number of Karate	4752	438	78
Edge number of Dolphin	10176	1835	159
Edge number of Football	39232	36543	631

It can be seen that the improved K-symmetric anonymity algorithm reduces many newly added nodes and edges compared to the original K-symmetric anonymity algorithm, greatly reduces the cost of anonymity. With the increasing of symmetrical replications, the number of nodes and edges in the networks and the costs of anonymity will also reduce greatly. And the results of the new restoration algorithm are also consistent with the original social networks, indicating that NI-PP is highly available.

VI. CONCLUSION

This paper designs a privacy protection method for social networks based on node importance, which includes a new restoration algorithm. In this method, the importance of nodes is further considered on the basis of K-symmetric anonymity. K-shell algorithm is used to screen out important nodes and protect them emphatically. The new restoration algorithm solves the problem that the ordinary restoration algorithm cannot completely restore the original structures, improves the availability of K-symmetric anonymity, and reduces the loss of information in the restoration process. NI-PP can protect social networks against the attacks based on structural knowledge.

ACKNOWLEDGMENT

This work is supported by the National Key R&D Program of China (2017YFB1401302, 2017YFB0202200), the National Natural Science Foundation of P. R. China (No.61572260, 61872196), Key R&D Program of Jiangsu (BE2017166).

REFERENCES

[1] Liu. Xiangyu and Wang. Bin, “Survey on Privacy Preserving Techniques for Publishing Social Network Data, ” Journal of Software . vol. 25(3), pp. 576-590, 2014.

[2] Lan. Lihui and Ju. Shiguang, “Social Networks Data Publication Based on K-anonymity,” Computer Science .vol. 38(11), pp.156-160, 2011.

[3] Jiang. Chen, Yang. Geng, Bai. Yunlu, and Ma. Junmei, “Frequent Itemsets Mining Algorithm for Privacy Protection,” Netinfo Security. pp. 73-81, 2019.

[4] L. Backstrom, C. Dwork, and J. Kleinberg, “Wherefore art thou R3579X? anonymized social networks hidden patterns and structural steganography, ”[International Conference on World Wide Web, pp. 181-190, 2007].

[5] M, HAY, G. MIKLAU, and D. JENSEN, “Resisting structural re-identifleation in anonymized social networks, ” Proceeding of the VLDB Endowment. vol.1(1), pp.102-114 , 2008.

[6] Zhou. Bin and Pei. Jian, “Preserving privacy in soda : networks against neighborhood attacks, ” [Proc of the 24th International Conference on Data Engineering, pp.506-515 , 2008].

[7] Y. Sei, H. Okumura, and T. Takenouchi, “Anonymization of sensitive Quasi-Identifiers for l-diversity and t-closeness, ”[IEEE Transactions on Dependable & Secure Computing, pp. 99, 2017].

[8] Zhu. Xiaoxia and Hu. Xiaoxue, “Identification of Node Influence Based on Improved k-shell Algorithm,” Computer Engineering and Applications .vol. 55(1), pp.33-41, 2019.