On Study of Privacy Preserving in Large-scale Social Networks Based on Heuristic Analysis

To cite this article: Huanxiang Zhang et al 2018 J. Phys.: Conf. Ser. 1087 062002

View the article online for updates and enhancements.
On Study of Privacy Preserving in Large-scale Social Networks Based on Heuristic Analysis

Huanxiang Zhang¹*, Xiaolin Zhang, Lixin Liu, Jing Zhang

School of Information Engineering, Inner Mongolia University of Science and Technology Baotou 014010, China
zhanghuanxiang@imust.cn

Email address: zhanghuanxiang@imust.cn.

Abstract. The social network structure refers abstract demonstration truly existed in real life and equipped with highly complicated networks; the social networks have been widely infiltrated into all aspects of people's daily life, such as the application of microblog, We Chat and others. With the rapid development of social networks, an explosive development trend occurs in the networks' scale whose data information is both various in types and rich in content, however, one has to notice that these data are related to people's privacy more or less at the same time, therefore people attach great importance to it and it is also essentially important to take measures for privacy preserving. This paper will carry out profound study on privacy preserving in large-scale social networks based on heuristic analysis.

1. Introduction
In recent years, more and more online social networks are built gradually and they get a large number of user groups; it is inevitable for users to leak privacy information when applying social networks, while with frequent occurrence of events that users' privacy information is leaked through various social networks, people attach great importance to studies on privacy preserving of social networks data. Therefore, the paper proposes to preserve privacy in large-scale social networks on the basis of heuristic analysis and hopes to provide referential values for relevant researchers.

2. Overview
Currently, social networks have penetrated into all aspects of people's daily life and play an increasingly critical role and value. The research performed by virtue of social networks not only fully demonstrates social networks' characteristics equipped by individuals but also promotes the understanding of network phenomenon in real society [1]. In view of the diversity of social networks, it is necessary to analyze social networks before and after the privacy preserving. The main duty of social networks is to carry out rigorous logical analysis of entities relevance in terms of relationships statement between people, institutions or other both sides with attributes of entities, as well as their value-evaluated process, and it can precisely interpret large-scale social networks so as to support gist for newly developed network system frameworks and models, and to develop new applications based on preserving of network privacy.
3. Heuristic Analysis

In current traditional social network frameworks, privacy preserving refers to a privacy attack way aiming at a single type, and however, it lacks comprehensive integrity and fails to cope with simultaneously multi-level privacy attacks. The study proposes heuristic analysis and, according to personalized privacy preserving of users, explores possibilities and strategies for carrying out network privacy preserving under the situation that current social network structures are basic stable so as to reduce possibilities and frequency that networks are attacked [2].

Currently, it is far from practical to identify defective amount of information through comparing social network information released before and after, meanwhile, the evaluation standard merely simply evaluates costs of privacy preserving rather than identifying which kind of information is defective[3]. Currently, in social network structures where image method is applied for description, it is not easy to consider many network framework factors including degree of connection, diameter and intermediate position of the image itself synchronously. Therefore, when designing network structures, one should find the balance between privacy preserving and information deficiencies as much as possible, and determine the network design goals: (1) to ensure that the released social networks will fully preserve privacy and prevent users’ privacy from leaking as much as possible; (2) to ensure that the social network structures shall maintain a relatively good stability so as to provide necessary data service support for carrying out network information application business.

It is necessary to design a new heuristic function because there may be conflicts among multi-objective functions which will fail to find optimal solution to optimize multi-objective in order to be close to the optimal solution of multi-objective functions infinitely [4].

1) In the actual design process, one should arrange corresponding preserving nodes according to personalized privacy preserving degree of users; privacy preserving degree of users at this node can be assumed as \(d_i\) which means the privacy preserving level of users, and then preserving degree of local network privacy can be accurately obtained:

\[P = \frac{\sum_{i} d_i}{|V|}\]

In the formula above, \(|V|\) refers the number of nodes whose privacy must be preserved in initial network.

2) The differentiated weight proportions can be laid according to objectives and requirements from users to network topological structure stability, as well as permissive errors of parameters like node degree distribution, network diameter and betweenness; meanwhile, the following formula can be adopted to calculate the network topological structure which changes before and after respectively, namely:

\[D = \sum w_i \times |a_i - a_i'|\]

In this formula, \(w_i\) represents weight proportions of structural parameter i, \(a_i\) represents the numerical calculation of structure parameter i in real network environment, and \(a_i'\) identifies the numerical calculation of structural parameter i in released network[5].

3) The heuristic function structure can be designed as follows:

\[F = \alpha \times P \times (1 - \alpha) \times D\]

where the range of \(\alpha\) is between 0–1. In this stage, the lower the function value, the better it will help to ensure the overall network can achieve better stability. Figure 1 below refers the whole flow chart of heuristic function.
Figure 1 Heuristic Function Flow

The statement 1 above shows that the key factor to realize heuristic function design is to minimize network topological differences before and after privacy preserving. The design of heuristic function mentioned above boasts of a good intuition, is easier to be achieved, can accurately be close to the optimal solution of objective function conforming true requirements and is conducive to strengthening the stability of social network topological structures.

4. Large-scale Social Network Privacy Preserving

With the rapid development of the current social networks, sizes of large number of social networks have reached ultra-large-scale data sets with ten million level, and it will certainly get difficult if privacy preserving approaches in traditional social networks are adopted based on the current situation; even though the privacy can be preserved, the preserving effect and performance of algorithm will certainly be significantly affected. Therefore it has become a major problem in analyzing social networks about how to effectively deal with large-scale social network data sets. Currently, the algorithm of mainstream social network privacy preserving just aims at ultra-large-scale social
networks and therefore, the paper studies relevant algorithms of fast privacy preserving of large-scale social networks, hoping to effectively cope with shortcomings in traditional privacy preserving algorithm and to promote the overall performance of algorithm.

4.1 Fast Community Mining
Currently, there are various community mining technologies in social networks, and therefore it is able to preserve privacy of large-scale community network rapidly according to the high efficiency and rapidness of algorithm itself and by applying fast community mining algorithm based on hierarchical clustering so as to comprehensively facilitate algorithm performance of overall privacy preserving; the construction thoughts of fast community mining algorithm based on hierarchical clustering are as follows:

(1) The community division modularity should be defined firstly, that is:

\[
\Delta Q_i = \frac{1}{2m} - \frac{k_i k_j}{(2m)^2}
\]

In the formula above, \(m\) represents the total number of edges in the network, and \(A_{vw}\) means whether there is a connection between the nodes V and W; \(k_v\) and \(k_w\) represent boundary values of nodes V and W respectively and \((c_v, c_w)\) is used by the Delta function to represent whether the communities where node V and W lie are same. In order to fully save space resources while accelerating the speed, it is possible to simplify the modularity \(Q\) to convert it into \(Q = \sum (e_{ij} - a_i^2)\) where \(e_{ij}\) and \(a_i^2\) respectively represent the number of connect edges of community i and j in total edges in the whole network, and they also represent overall ratio between number of edges connected with community i and the total number of edges [6].

(2) In the initial stage, each node in the network is in an independent community, and \(\Delta Q_i\) refers newly increased modularity after community i and j merger; the state of initial stage should be consistent with the following two requirements, namely, connect edge of community i and j is

\[
\Delta Q_{ij} = \frac{1}{2m} - \frac{k_i k_j}{(2m)^2} \text{ or } k_i
\]

The sparse matrix \(\Delta Q\) can be used to store \(\Delta Q_{ij}\) due to the merger of community i and j in social networks; the max heap H can be adopted to select the largest \(\Delta Q_{ij}\) for storage in \(\Delta Q\), and to store array of \(a_i\).

(3) The largest \(\Delta Q_{ij}\) and community i and j corresponding should be determined in max heap H. If the increments \(\Delta Q_{ij}\) of modularity is \(\geq 0\), community i and j can be merged.

The initial social networks can be separated into a series of community structures by applying fast community mining techniques mentioned above; in terms of internal nodes in each community structure relative to the overall network, the privacy preserving scales need to handling will be greatly reduced, and the performance of overall privacy preserving algorithm shall be comprehensively upgrade, besides, this algorithm performance can be close to linear growth trend infinitely. After that, the K-Degree anonymity can be applied to promote the K degree value of nodes in community to achieve anonymity so as to hidden topological structure information in each node.

4.2 Personalized K-Degree Anonymity
After the community mining is completed in original networks, the nodes in the original networks can
be divided into several community structures, and then personalized K-Degree anonymity can be applied to hide and protect nodes in the community; the personalized requirements hope that privacy preserving required by each object is different in networks and that proper measures should be taken to preserve privacy for a number of different objects. Considering the fact that illegal attackers often use a certain network technology to identify the target nodes in current situations, we find such technologies tend to be different to some extent. The structural information commonly used in current illegal attack behaviors is mainly node degree information which achieves a malicious attack to nodes at a time [7].

K-Degree anonymity is one of the most commonly used privacy preserving technologies in current network node anonymity technologies. Before the anonymity is implemented, the tag attributes of G node in original network should be removed firstly, which will produce a simple anonymous network whose privacy will be preserved specially in order to produce network G. In the whole process of individualizing K-Degree anonymity, it is necessary to arrange anonymous node in the community structure according to the degree level, and then summarize degree level information of each node so as to generate triad information ensemble that should be met by nodes anonymity. Once the number of nodes fails to reach K value in the case where K-Degree is waiting for an anonymous conversion, it is necessary to apply subsequent low-level nodes with same level of and make noise margin so as to ensure the number of passed nodes is not less than K. The $k_i$ level shall be obtained after triad ensemble of communication of all nodes is gotten, and the number of noise edges shall be newly increased through follow-up nodes.

Steps of quick community digging are: to define modularity $\rightarrow$ to initialize each node as an independent community $\rightarrow$ to join the community collection $\rightarrow$ to select two largest communities of modularity increment $\Delta Q$ $\rightarrow$ to merger the corresponding two communities [8].

4.3 Nodes Isomorphism

After the anonymous network processing is completed, the produced cohesion network scale will be significantly reduced, however, the illegal attackers can improve rate of identification of target nodes by relying on first order adjacent structure information of nodes and roughly identify specific scope of networks where target nodes exist so as supply privacy to target nodes; all in all, it is significantly important to hide attribute information of cohesion network nodes.

The network structure privacy attack can be defined as follows: suppose there is an unleashed network G, and if Q query of an illegal intruder can identify a limited number of nodes matching the target node t, then the target node t will be identified. If the Q query refers structural information based on the target node t, the attack can be identified as an attack on privacy. Currently, the structural privacy attacks commonly used are: sub-map structural attack, degree structural attacks, adjacent sub-map structural attacks and central node structural attack. Once measures are taken to preserve privacy of cohesion network after privacy attack behavior occurs, it is of little possibility that the target nodes will be attacked further. In this occasion, the node isomorphism can be used to preserve privacy of nodes in cohesion network. Node isomorphism refers that topological structures of adjacent nodes with same degree in network are same, which means there is a mutual mapping relation among them.

5. Conclusions

In a word, with the rapid development of social networks in recent years, the privacy preserving of users in social networks earns more and more attention from people. In order to effective solve the issue, the paper proposes the fast algorithm by applying a fast community mining algorithm to split the large-scale networks into subnets easy to carry out distributed privacy preserving so as to rapidly preserve privacy of large-scale social networks; meanwhile, the heuristic analysis is adopted to effectively reduce changes in network topological structures and to maximize privacy preserving required by users.
Acknowledgment
Project supported by National Natural Science Foundation of China (No.61562065) and Natural Science Foundation of Inner Mongolia (No. 2017MS(LH)0603) and The scientific research project of the Inner Mongolia Autonomous Region colleges and Universities(No. NJZY156).

References
[1] Gong Weihua, Lan Xuefeng and Pei Xiaobing. Privacy Preserving Method Based on k-Degree Anonymity in Social Networks[J]. Acta Electronica Sinica, 2016, 44 (6).
[2] Zhang Xiaolin, Guo Yanlei and Wang Jingyu. Privacy Preserving Method for Social Network Based on Pregel-like[J]. Application Research of Computers, 2016,33 (9)
[3] Li Jing and Han Jianmin. Privacy Preserving Method for a Kind of Sensitive Relation Social Network - (k, 1) - Anonymous Model [J]. Journal of Chinese Computer Systems, 2013,34 (5)
[4] Lan Lihui and Ju Shiguang. Privacy Preserving Method Based on Random Projection for Weighted Social Networks[J]. Computer Science, 2016, 43 (3)
[5] Zhang Wei, Wang Xuran and Wang Yu. Privacy Preserving Method of Dynamic Social Network Based on K- neighborhood Isomorphism[J]. Journal of Nanjing University of Posts and Telecommunications (natural science edition), 2014, 34 (5)
[6] Zhang Xiaolin, Zhang Chen and Zhang Wenchao. D-VSSP: Privacy Preserving Algorithm of Distributed Social Network[J]. Computer Science, 2017, 44 (2)
[7] Liu Xiangyu, Wang Bin and Yang Xiaochun. Overview of Privacy Preserving Technologies for Social Network Data Publishing[J]. Journal of Software, 2014, 25 (3)
[8] Lan Lihui and Ju Shiguang. Privacy Preserving Based on Differential Privacy for Weighted Social Networks[J]. Journal of Communications, 2015, 36 (9).