Privacy Protection in Online Social Network in the Context of Big Data

Pingshui WANG[a,*], Zecheng WANG[b], Tao CHEN[c]

College of Management Science and Engineering, Anhui University of Finance & Economics, Bengbu 233040, China

[a]email: pshwang@163.com, [b]email: zcwang@ah.edu.cn, [c] echentao@163.com, [*]corresponding author

Keywords: privacy protection; online social network; big data

Abstract: With the rapid development of Web 2.0 technology, online social network data has shown the classical big data characteristics. The analysis of social network big data is getting deeper and deeper. At the same time, the risk of privacy disclosure in social network is also very obvious. A series of privacy preserving models and algorithms for social network data are proposed. In this paper, we summarized the privacy leakage type of social network, deeply analyzed the existing privacy preserving technology, pointed out its advantages and disadvantages, and prospected the future research directions.

Introduction

With the rapid development of Internet and information technology, especially the application of Web 2.0 technology, online social networks (OSNs) have experienced exponential growth in recent years. All kinds of social network products were introduced to the Internet, such as Facebook, Myspace, Twitter, RenRen, WeChat, etc. OSNs have already become the new media platform with the most users and the biggest impact. OSNs can easily collect a large number of user data, which are the classical big data for its 4V features: the large scale amount data (Volume), complex data types (Variety), very fast growth (Velocity) and considerable data availability (Veracity). The vigorous development of big data brought opportunities and challenges to social networks, social networks in the era of big data presents new characteristics. OSNs provide much convenience for human mutual communication, information sharing and data analysis. At the same time, it also brings great threat and challenge to individuals' privacy, because OSNs data may contain personal private information. Protecting the privacy of users against unwanted disclosure in such circumstance poses challenging problems. Issues on privacy disclosure are the greatest threat to the personal information security in the era of big data, and the biggest obstacle to the development of big data.

In recent years, the issues on privacy protection of relational data release are deeply studied, and lots of effective privacy preserving technologies have been developed, which mainly include k-anonymity, l-diversity, t-closeness, and m-invariance, and so on. However, these relational data privacy preserving techniques cannot be directly applied to social network data privacy protection. Because relational data privacy preserving model considers each record is fully independent to all other records, ignoring the relationship between nodes in the social network, which can also be considered as the attacker's background knowledge [1]. The social networks are usually modeled as graph, in which social roles are nodes and social relationships are edges. Therefore, relational data privacy preserving technology cannot meet the social network data privacy preserving requirements. There is relatively less research work on privacy protection of social network data.

In this paper, we summarized the privacy leakage type of social network, deeply analyzed the existing privacy preserving technologies [2-18] in social network, pointed out their advantages and disadvantages, and prospected the future research directions.

Privacy Attacks in Social Networks

Social network data mainly includes nodes representing entities, edges representing the relationship between entities, and graph structure formed by nodes and edges, any of which may contain private information, and may be attacked. The following is a brief introduction to the attacks related to social network nodes, edges, and structural information.

- **Attacks on Social Network Nodes**

Social network nodes represent a kind of social role, it may be an individual, a group or an organization, any information contained in a node may become attacked object, such as its attribute information, existence and structure.

Each node in the social network has its attribute value, some of which may be related to personal privacy, such as personal income, medical records, personal preference, religious belief, etc. When social network data is published, the mutual relationship between the nodes makes the attacker have more background knowledge to speculate the sensitive attribute information of the target node.

The existence of node is whether an individual is in a social network in the form of a node. In some cases, some people regard his/her appearance in a certain social network as privacy. Therefore, the information should be prevented from the combination of background knowledge.

Not only some attribute values of the nodes are sensitive, but also the structural properties of nodes in the social network are considered as privacy, such as the degree of nodes, the shortest distance between two nodes, the distance between the nodes and the center of the social network.

- **Attacks on Social Network Edges**

Social network edges present the relationship between various social roles. It may be friends, colleagues, leadership, and cooperation and so on. Similar to nodes, the attribute, the existence, as well as the weight of the edge may become attacked target.

Similar to node attribute values, edges of the social network can also have attribute value, such as the label of the edge can be expressed as the relationship type between the nodes.

The so-called edge existence refers to whether two designated nodes in the social network have a certain relationship. If an edge between the two nodes is sensitive, deleting the edge of the two target nodes cannot protect the private information, because the attacker can also speculate whether the edge is sensitive through the background knowledge.

Edge may have weight in specific application, when edge’s weight is regarded as private information, it should also be protected.

Privacy Preserving Technology in Social Networks

The present social network privacy preserving technology is mainly divided into 3 types: node K-anonymity, subgraph K-anonymity and data disturbance.

The main idea of the node K-anonymity [2-6] and subgraph K-anonymity [7-13]: there are at least k candidates in the anonymized social network while attackers identify a special object based on background knowledge, that is to say, the probability of privacy disclosure is less than 1/K.

The main idea of data disturbance is that the social network data is randomly changed, which makes the attacker cannot accurately infer the original real data. Data disturbance method is divided into value disturbance [14-16] and graph structure disturbance [4,18].

- **Node K- anonymity**

The so-called node K- anonymity, refers to all nodes in the social network are clustered into a number of super points, each of which contains at least K nodes. Due to the nodes in the super points cannot be distinguished from each other, so the privacy leakage probability of node re-identified attack is less than 1/K. However, node clustering has resulted in information loss of the clustered nodes, increased the uncertainty of the graph structure, so the availability of the released anonymous graph data is reduced.

The paper [2] studies how to minimize the possible social network count |W(G)| while achieving node K- anonymity based on the idea of simulated annealing. Based on literature [2], the improvement is made in [3], which assumes that each node in the social network has attribute information. A greedy clustering method is proposed to implement the node K- anonymity for complex social network. But this anonymity algorithm requires the data publisher focus on reducing the loss of graph structure information or loss of node attributes through setting the weight, as the data availability is difficult to quantify, the weights cannot be properly set, so this method is very poor in practical application.

Node K-anonymity technique has high privacy preserving ability and good generality. However, for the high information loss and low efficiency, node K-anonymity mothed is not suitable for large scale social network data.

- **Subgraph K-anonymity**

Subgraph K-anonymity refers to when the attacker want to access the private information regarding the particular subgraphs consisting of target node as background knowledge, there are at least K candidate subgraphs in the social network, so the probability of subgraph privacy leakage is less than 1/K. Subgraph K-anonymity can be generated by adding pseudo point, pseudo edge, edge deletion, generalization, suppression and so on.

Paper [7] studies how to achieve subgraph K-anonymity through the degree of node as background knowledge, and proposes K-degree anonymity algorithm using dynamic programming method so as to the number of added edges is the least. In literature [9], the loss of community structure information is minimized while the K-degree anonymous graph is generated. In paper [10], K-automorphism method is proposed to protect privacy. K-automorphism means the graph itself has a K-automorphic mapping. This method can prevent node re-identification attack, but cannot prevent sensitive relationship attack. To protect node and edge privacy, paper [11] proposes K-isomorphism privacy preserving model, which means the social network graph is divided into K subgraphs, and these subgraphs are isomorphic. As the edge between the isomorphic subgraphs is removed, the graph pattern would be affected.

- **Data Disturbance**

The basic idea of the graph data disturbance method is that a random change is made in social network graph, which makes the attacker cannot accurately guess the original real data, so as to protect the social network data privacy. In the following, we introduce the social network privacy preserving technology from the aspects of value disturbance and graph structure disturbance.

In literature [14], the influence on the shortest path sequence and the shortest path of two nodes in social networks resulting from reducing disturb noise is studied by disturbing technology to protect the privacy of the social network edge weight. In [15], a linear programming model is proposed to maintain the linear graph property while the weight of an edge is disturbed, which can guarantee the availability of the weighted graph. In [17], an edge vector perturbation method to preserve structural properties and edge weights for weighted social networks is proposed, which can be applied to a typical perturbation algorithm to achieve better preservation of the utility of its output. To maintain the availability, literature [18] studied how to maintain the same graph spectrum while disturbing the graph, which points out that the graph spectrum is determined by two parameters: the adjacency matrix of graphs with the largest eigenvalue and the secondary minimum feature values of Laplacian matrix.

Future Work

Privacy protection in social network is a new research direction. There are still many problems that should be further studied.

- **Research on dynamic privacy preserving technology**

The current privacy preserving techniques are only used to static social network data release, but social network data are dynamic changed from time to time. Therefore, it is necessary to develop dynamic privacy preserving technology. Otherwise, due to the inference channel existing between two released versions, the published social network data would be threatened by preforming inference attack.

- **Research on parallel privacy preserving technology**

The present social network analysis and privacy preserving technologies are not suitable for social network big data. Therefore, it is necessary to study the parallel processing technology. For a social network privacy preserving algorithm, the more important is the efficiency and ability to expand to large-scale data.

- **Research on personalized privacy preserving technology**

To meet the time challenge of online social network in the context of big data, in view of the problems of high information loss and low data availability caused by existing privacy preserving scheme for its "over-protection", personalized social network privacy preserving model and algorithm should be developed, which allows users to express personalized privacy preferences.

- **Research on new privacy preserving technology supporting specific data application**

The current social network privacy preserving research has not specified the use of published data, it only design a general privacy preserving method, which affects the availability of published data. Therefore, it is necessary to implement privacy protection based on the use of published data, so as to improve the availability.

Summary

In recent years, privacy preservation has been widely concerned in academic and industrial fields. Many privacy preservation techniques have been proposed. On the basis of full investigation and deep analysis, this paper summarizes the research progress of privacy protection technology in online social network. The privacy leakage type of social network data is list, the current privacy protection technologies are deeply analyzed, their advantages and disadvantages are also pointed out and the future research directions are prospected.

Acknowledgments

We thank the anonymous reviewers and editors for their very constructive comments. This work was partly supported by the Anhui Provincial Institutions of Higher Learning Natural Science Project of China (No. KJ2015A106, KJ2013A001).

References

[1] Liu XY, Wang B, Yang XC. Survey on privacy preserving techniques for publishing social network data [J]. Journal of Software, 2014,25(3):576-590.

[2] Hay M, Miklau G, Jensen D, Towsley D. Resisting structural identification in anonymized social networks [C]. Proceedings of the 34th Int'l Conf. on Very Large Databases. 2008,102-114.

[3] Campan A, Truta TM. A clustering approach for data and structural anonymity in social networks [C]. Proceedings of the 2nd ACM SIGKDD Workshop on Privacy, Security, and Trust in KDD. 2008,33-54.

[4] Cormode G, Srivastava D, Yu T, Zhang Q. Anonymizing bipartite graph data using safe groupings [C]. Proceedings of the 34nd Int'l Conf. on Very Large Databases. 2008,833-844.

[5] Tai C, Yu P, Yang D, Chen M. Privacy-preserving social network publication against friendship attacks [C]. Proceedings of the SIGKDD 2011,1262-1270

[6] Wang R, Zhang M, Feng D, Fu Y. A clustering approach for privacy-preserving in social networks [J]. Lecture Notes in Computer Science 8949, 2014,193-204.

[7] Liu K, Terzi E. Towards identity anonymization on graphs [C]. Proceedings of the 2008 ACM SIGMOD Int'l Conf. on Management of Data. 2008, 93-106.

[8] Yuan M, Chen L, Yu PS. Personalized privacy protection in social networks [C]. Proceedings of the 36th Int'l Conf. on Very Large Databases. 2010,141-150.

[9] Wang Y, Xie L, Zheng B, Lee KCK. Utitily-Oritented k-anonymization on social networks [C]. Proceedings of the 16th Int'l Conf. on Database Systems for Advanced Applications. 2011,78-92.

[10] Zou L, Chen L, Ozsu M. K-automorphism: a general framework for privacy preserving network publication [C]. Proceedings of the the 35th Int'l Conf. on Very Large Databases, 2009,2(1):946-957.

[11] Cheng J, Fu AWC, Liu J. K-isomorphism: Privacy preserving network publication against structural attacks [C]. Proceedings of the 2010 ACM SIGMOD Int'l Conf. on Management of Data. 2010,459-470.

[12] Campan A, Truta T. Data and Structural k-Anonymity in Social Networks [C]. Proceedings of the PinKDD 2008. LNCS 5456, 2009,33–54.

[13] Sun C, Yu P S, Kong X, Fu Y. Privacy preserving social network publication against mutual friend attacks [J]. Transactions on Data Privacy, 2014, 7(2):71-97.

[14] Liu L, Wang J, Liu J, Zhang J. Privacy preserving in social networks against sensitive edge disclosure [R]. Technical Report, CMIDAHiPSCCS006-08, Department of Computer Science, University of Kentucky, 2008.

[15] Das S, Egecioglu O, Abbadi A. Anonymizing weighted social network graphs [C]. Proceedings the 26th Int'l Conf. on Data Engineering,2010,904-907.

[16] Li Y, Shen H. Anonymizing graphs against weight-based attacks [C]. Proceedings of the ICDM Workshops 2010,491-498.

[17] Lan L, Tian L. Preserving social network privacy using edge vector perturbation [C]. Proceedings of the International Conference on Information Science and Cloud Computing Companion, 2014,188-193.

[18] Ying X, Wu X. Randomizing social networks: A spectrum preserving approach [C]. Proceedings of the 2008 SIAM Int'l Conf. on Data Mining, 2008,739-750.