

Privacy Preservation in Social Network Based on Anonymization Techniques

Pingshui Wang^{1*}, Xuedong Zhang¹, Pei Huang²

¹College of Management Science and Engineering, Anhui University of Finance & Economics, Bengbu 233030, China;

²College of Computer Science and Technology, Jilin University, Changchun 130012, China

Received 1 October 2014, www.cmnt.lv

Abstract

Social network data can be released for various purposes, such as statistical analysis, cluster queries, data mining and so on. However, when social network data are released, the privacy of some individuals and organizations may be disclosed. Serious concerns on privacy preservation in social networks have been raised in recent years. In this paper, in order to reduce the privacy disclosure, we propose a privacy preservation model and algorithm based on anonymization techniques for social network data, that is (k, l) -anonymity algorithm. The proposed (k, l) -anonymity algorithm anonymize social network data to prevent privacy attacks including both content and structural information, while minimizing the anonymization cost and reducing the privacy disclosure. Extensive experiments have been conducted on synthetic data sets comparing with previous work. The result shows that the proposed anonymity algorithm could improve the security of the released social network data while maintaining data utility.

Keywords: Social network, privacy preservation, anonymization, privacy disclosure risk.

1 Introduction

With the rapid development of Internet, online social networks (OSNs) have experienced exponential growth in recent years, for example, Facebook, Myspace, Twitter and so on. A large part of the growth is owing to the ease of sharing information with other users who have common interests. As a result, OSNs store a huge amount of possibly sensitive and private information on users and their interactions. In order to mine the potential value of the social network data, many analysis methods have been proposed. Preserving the privacy of individuals against unwanted disclosure of their information in such circumstance poses serious challenging privacy issues.

In the recent years, a number of effective methods for protecting privacy of relational data have been proposed. The application is mainly data releasing so that sensitive individual information can be protected while organizations are releasing their data such as credit records, medical records, transactions records, and so on. These methods include k -anonymity model [1,2], l -diversity model [3], t -closeness model [4], and m -invariance model [5]. A naive technique of preserving privacy is deleting attributes that uniquely identifying an individual such as names and identification card numbers. However, trivial linking attack by using the innocuous sets of attributes named as quasi-identifiers across multiple databases can easily identify an individual. K -anonymity model [1,2] is the first attempt of privacy preservation of relational data by ensuring at least k records with respect to every set of quasi-identifier attributes are indistinguishable. However, k -anonymity is

reluctant when there is a lack of diversity of sensitive attribute value or other background knowledge. L -diversity model [3] ensures that there are at least l well-represented values of the sensitive attributes for every set of quasi-identifier attributes. The weakness is that one can still estimate the probability of a particular sensitive attribute value. M -invariance model [5] ensures that each set of quasi-identifier attributes has at least m tuples, and each with a unique set of sensitive attribute values. There is at most $1/m$ confidence in identifying the sensitive values. However, these methods in privacy preservation of relational data are not directly applicable in social network data because the data representations are different. The challenge is to develop techniques to release social network data in a pattern that preserves utility without compromising privacy. There is relatively less research work on privacy preservation of social network data. Previous work has proposed various privacy models and algorithms with the corresponding preservation mechanisms that prevent both inadvertent private information disclosure and attacks by malicious adversaries. Those early privacy models are mostly related to content and structural information disclosure. The social networks are modeled as graphs in which users are nodes and social connections are edges. The threat definitions and preservation mechanisms leverage structural properties of the graph. Facing with the problem, several works [6-13] develop methods that can be applied to the simple anonymized graph, further modifying the graph in order to provide comprehensive privacy guarantee. Some works are based on graph models other than native graph [14-16].

*Corresponding author's e-mail: pshwang@163.com

To our knowledge, Zhou et al. [17] and Yuan et al. [18] were the first to consider modeling social networks as labeled graphs, similarly to what we consider in this paper. The method to achieve the above anonymities is edge or node distortion. By adding and/or deleting edges and/or nodes, a distorted graph is generated to satisfy the anonymity requirement. Attackers could only have a probability of $1/k$ to recognize the identity of a node by neighborhood attacks. To prevent re-identification attacks by adversaries with direct neighborhood structural knowledge, Zhou et al. [17] propose a technique that groups nodes and anonymizes the neighborhoods of nodes in the same group by generalizing and/or suppressing node labels and adding and/or deleting edges. They enforce a k -anonymity privacy restriction on the graph, each node of which is ensured to have the same direct neighborhood structure with other $k-1$ nodes. Yuan et al. [18] try to be more practical by considering individuals' different privacy concerns. They divide privacy restrictions into three levels, and propose techniques to generalize labels and modify structure corresponding to each privacy requirement. Nevertheless, both Zhou and Yuan have not considered labels as a part of the background knowledge. However, in case attackers hold label information, these methods can not achieve the same privacy insurance.

In this paper, we discussed privacy risks in releasing social network data and the design principles for improving security and reduce anonymization cost. Aiming at the content disclosure and structure disclosure of the previous work, we proposed an effective privacy preserving algorithm: (k, l) -anonymity algorithm for social network. Extensive experiments have been conducted on synthetic data sets. The result shows that the proposed anonymity algorithm could improve the security of the released social network data while maintaining data utility.

The rest of this paper is organized as follows. In section 2, we introduce the related concepts. In section 3 and section 4, we present our (k, l) -anonymity model and algorithm for social network data publishing respectively. In section 5, we analyze the performance of our technique via extensive experiments. Section 6 contains the conclusions and future work. Section 7 declares the conflict of interest. Acknowledgments are given in section 8.

2 The Related Concepts

In this section we introduce two models that will be used in the paper.

2.1 K-ANONYMITY MODEL

In order to preserve privacy, Sweeney et al. [2] proposed the k -anonymity model, which achieves k -anonymity using generalization and suppression techniques, so that, any individual is indistinguishable from at least $k-1$ other ones with respect to the quasi-identifier attributes in the released dataset. For example, table 2 is a 2-anonymous table of Table 1.

TABLE 1 Original table.

Name	Race	Birth	Sex	Zip	Disease
Alice	Black	1965-3-18	F	02141	gastric ulcer
Helen	Black	1965-5-1	F	02142	dyspepsia
David	Black	1966-6-10	M	02135	pneumonia
Bob	Black	1966-7-15	M	02137	bronchitis
Jane	White	1968-3-20	F	02139	flu
Paul	White	1968-4-1	F	02138	cancer

TABLE 2 2-Anonymization of table 1.

Race	Birth	Sex	Zip	Disease
Black	1965	F	0214*	gastric ulcer
Black	1965	F	0214*	dyspepsia
Black	1966	M	0213*	pneumonia
Black	1966	M	0213*	bronchitis
White	1968	F	0213*	flu
White	1968	F	0213*	cancer

2.2 L-DIVERSITY MODEL

Since k -anonymity algorithm is reluctant to background knowledge attack and homogeneity attack, Machanavajjhala et al. [3] proposed l -diverse anonymity model, which requires each equivalence class contain at least l "well-represented" values for the sensitive attribute. As a simple and direct interpretation, l -diverse anonymity means that each equivalence class contains at least l different sensitive attribute values. For example, table 2 is also a 2-diverse table of table 1.

3 (k, l) -Anonymity Model For Social Network

In this paper, we model a social network as a simple graph $G = (V, E)$, where V is the set of vertices corresponding to the individuals, and $E \subseteq V \times V$ is the set of edges representing the relationships between the individuals.

Based on the above observation, we define the anonymization cost incurred by edge addition and deletion.

Definition 1. Anonymization Cost. Given a weight ω , $0 < \omega \leq 1$, the cost of anonymizing $G = (V, E)$ to $\bar{G} = (V, \bar{E})$ is

$$Cost(G, \bar{G}) = \omega |\bar{E} / E| + (1 - \omega) |E / \bar{E}|.$$

When ω is close to zero, we prefer to preserve the existing edges and generate a super graph \bar{G} of G . In contrast, a larger ω results in more edge deletions.

Definition 2. k -Isomorphic Graph ($k-IG$). A graph $G = (V, E)$ is k -isomorphic, if G consists of k disjoint subgraphs g_1, g_2, \dots, g_k , i.e. $G = \{g_1, g_2, \dots, g_k\}$, where g_i and g_j are isomorphic if $i \neq j$.

Definition 3. k -Isomorphic Vertex Set ($k-IVS$). Given a k -isomorphic graph $G = (V, E)$, the set of vertex v_i and v_j is an k -isomorphic vertex set, if for $\forall v_i \in V$, $\exists k-1$ vertexes $v_j \in \bar{V}(j \neq i)$ are isomorphic to v_i .

Similar to definition 2, we define k -isomorphic edge set in the following.

Definition 4. k -Isomorphic Edge Set ($k-IES$). Given a k -isomorphic graph $G = (V, E)$, the set of edges e_i and e_j is an k -isomorphic edge set, if for $\forall e_i \in E$, $\exists k-1$ edges $e_j \in \bar{E}(j \neq i)$ are isomorphic to e_i .

Definition 5. (k, l) -Anonymous Graph. Given a k -isomorphic graph $G = (V, E)$, the graph \bar{G} is (k, l) -

anonymous graph, if all the k -isomorphic vertex set and k -isomorphic edge set are l -diverse.

Based on the above definitions, we define the anonymization problem considered in this paper as follows.

Definition 6. (k, l)-Anonymous Problem. Given an undirected graph $G = (V, E)$, a positive integer k , and a sensitive parameter l , the graph $G = (V, E)$ is an anonymous graph of $G = (V, E)$, the problem is to minimize $Cost(G, \bar{G})$ for anonymizing $G = (V, E)$ to $\bar{G} = (V, E)$ such that $\bar{G} = (V, E)$ is (k, l) -anonymous.

To solve the problem, we present an effective heuristic algorithm for anonymizing large-scale social networks, details in section 4. To limit the distortion in $G = (V, E)$, we preserve the vertex set, i.e., $V = V$, and allow only edge addition operations. Preserving the vertex set prevents the addition or deletion of any individual.

4 (k, l)-Anonymity Algorithm For Social Network

Consider a social network $G = (V, E)$, a positive integer k , and a sensitive parameter l , our algorithm follows these rules. In anonymizing process, no fake vertex is added. This constraint is desirable, because the fake vertices usually change the whole structure of social networks. Moreover, we modify the graph only by adding edge, not by deleting the edge.

The objective of (k, l) -anonymity of $G = (V, E)$ is to produce a releasable anonymous graph $\bar{G} = (V, E)$ such that (i) $\bar{G} = (V, E)$ is (k, l) -anonymous, (ii) the $Cost(G, \bar{G})$ is minimal. However, the problem of optimal (k, l) -anonymity of $G = (V, E)$ is NP-hard. Therefore, we develop a heuristic (k, l) -anonymity algorithm.

Our heuristic (k, l) -anonymity algorithm includes three steps. The basic idea is in the following.

Given a social network graph $G = (V, E)$, firstly, we derive a k -isomorphic graph $G' = (V, E')$ by adding edges such that it is k -isomorphic and the $Cost(G, G')$ is minimal.

Secondly, to generalize the quasi-identifier attributes of all the k -isomorphic vertex set of $G' = (V, E')$.

Finally, to generate a releasable anonymous graph $\bar{G} = (V, E)$ by generalizing the quasi-identifier attributes of all the k -isomorphic edge set of $G' = (V, E')$ so that $Cost(G, \bar{G})$ is minimal.

Hence, the quality of the releasable anonymous graph $\bar{G} = (V, E)$ depends on (i) the method of generalization, and (ii) the choice of the local recoding techniques.

Algorithm:

(k, l) -Anonymity Algorithm for Social Network

Input: an original social network graph $G = (V, E)$, an anonymous parameter k , and a diverse anonymity threshold value l .

Output: the releasable graph $\bar{G} = (V, E)$, which is (k, l) -anonymous and the $Cost(G, \bar{G})$ is minimal.

1. For the given graph $G = (V, E)$, under the condition of minimal $Cost(G, G')$, in accordance with the anonymous parameter k , derive a graph $G' = (V, E')$ by adding edges such that G' is k -isomorphic.
2. Compute all the k -isomorphic vertex set $k-IVS$ and all the k -isomorphic edge set $k-IES$ of $G' = (V, E')$.
3. Compute the count of the $k-IVS$ and $k-IES$: $|k-IVS|$ and $|k-IES|$.
4. For each $k-IVS$
 - 1) generalize every categorical quasi-identifier attribute of the $k-IVS$ to a more general value in accordance with its taxonomy.
 - 2) generalize every numeric quasi-identifier attribute of the $k-IVS$ to an interval.
 - 3) generalize sensitive attribute till all the $k-IVS$ satisfying l -diversity.
5. For each $k-IES$
 - 1) generalize every categorical quasi-identifier attribute to a more general value in accordance with its taxonomy.
 - 2) generalize every numeric quasi-identifier attribute to an interval.
 - 3) generalize sensitive attribute till all the $k-IES$ satisfying l -diversity.
6. Return a releasable anonymous graph $\bar{G} = (V, E)$.

5 Experimental Results

The main goal of the experiments was to investigate the performance of our algorithm in terms of risk of privacy disclosure, data utility and implementation efficiency. To accurately evaluate our algorithm, we compared our implementation with the previous work that is the k -degree anonymity algorithm proposed by Liu and Terzi [9].

5.1. EXPERIMENTAL SETUP

In our experiments, we adopted synthetic data sets: using the R-Mat graph model [19] to generate synthetic graphs. The model takes four parameters a, b, c and d as inputs, where $a + b + c + d = 1$. The generated graphs have the power-law degree distributions and small-world properties, which are observed in many real world social networks. In this work, we use the suggested settings (0.45, 0.15, 0.15 and 0.25) for the four corresponding parameters, and generate several different size of social network data sets, which consists of 10,000 to 50,000 vertices and 80,000 to 400,000 edges. Each vertex is assigned a group of attribute values, including $\{age, gender, race, marital status, native country, occupation, disease\}$. We considered $\{age, gender, race, marital status, native country, occupation\}$ as quasi-identifier, $disease$ as sensitive attribute, whose values are randomly generated and satisfy Gauss distribution, as shown in Table 3. Similar to vertices, each edge is also randomly assigned a group of attribute values.

The experiments were performed on a machine with Intel(R) Core(TM)2 Duo CPU T5450 1.67GHz(Double Kernel), 4.0GB RAM, Windows XP, MATLAB7.0, and Visual C++ 6.0.

TABLE 3 Attribute information

	Attribute	Distinct Values	Sensitive?
1	age	74	No
2	gender	2	No
3	race	5	No
4	marital status	7	No
5	native country	41	No
6	occupation	14	No
7	disease	12	Yes

5.2. RISK OF PRIVACY DISCLOSURE

The risk of privacy disclosure can be defined as the probability of inferring privacy by the attacker according to the social network graph and related background knowledge. Let s be sensitive data, S_k be the event “the attacker exposes sensitive data based on background knowledge”. The risk of privacy disclosure $r(s, K)$ can be defined as:

$$r(s, K) = \Pr(S_k)$$

If the risk of privacy disclosure of all sensitive data of a released social network graph is less than a threshold α ($0 \leq \alpha \leq 1$), we name the $r(s, K)$ of the social network graph as α .

Figure 1 compares the risk of privacy disclosure of the two anonymous social network graph generated by different anonymity algorithms. As the figure illustrating, our algorithm results in lower risk of privacy disclosure than the k -degree anonymity algorithm for all the given k value ranging from 50 to 500. That is because we consider the sensitive data diversity of vertexes and edges except for the anonymity.

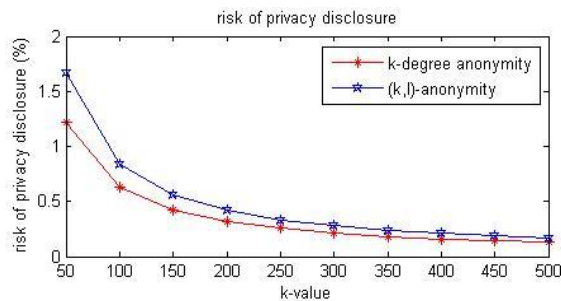


FIGURE 1 Risk of privacy disclosure.

5.3 DATA QUALITY

In this section, we report experimental results on our algorithm and the k -degree anonymity algorithm for data quality.

We test the data quality by executing aggregation queries on the released anonymous social network graph. Vertex pairs are randomly selected for testing the average shortest path between Vertexes. The query error rate of average shortest path is calculated as follows:

$$r = \frac{|d - d'|}{d}$$

Where d and d' were average shortest path length of the original social network and anonymous social network.

Figure 2 shows the query error rate of average shortest path of the two algorithms for the increasing size of social network graph. As the figure illustrating, our algorithm results in little more query error rate than the k -degree anonymity algorithm for all size.

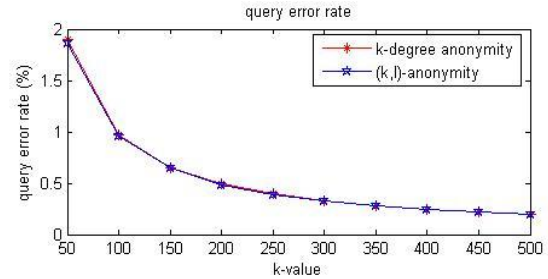


FIGURE 2 Data quality.

5.3. IMPLEMENTATION EFFICIENCY

The execution time of the two algorithms is shown in Figure 3. The execution time of our algorithm is always greater than that of the k -degree anonymity algorithm. The reason is that, our algorithm needs to preserve the l -diversity of the k -isomorphic vertex set and k -isomorphic edge set except for considering the property of k -anonymity and k -isomorphism, which guarantees the released social network less risk of privacy disclosure.

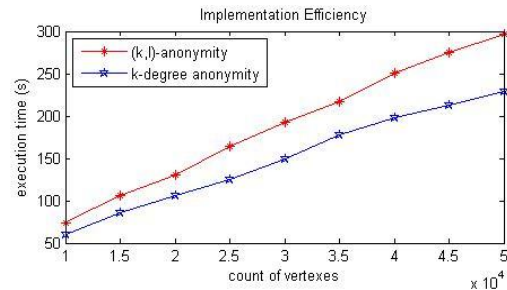


FIGURE 3 Implementation efficiency.

6 Conclusions and Future Work

In recent years, privacy preservation has been widely concerned in academic and industrial fields. Many kinds of privacy preserving methods have been proposed. However, there is relatively less research work on privacy preservation of social network data. In this paper, aiming at the higher risk of privacy disclosure, on the basis of the previous work, we studied the issues on social network data releasing and proposed an effective privacy preserving algorithm, that is (k, l) -anonymity algorithm for social network. Through extensive experiments, we verify the effectiveness and applicability of the proposed technique. In the future, we will further research and develop more effective and efficient anonymity algorithm for privacy preserving social network data release.

7 Conflict of Interest




We declare that this article content has no conflict of interest.

Acknowledgments

We thank the anonymous reviewers and editors for their very constructive comments. This work was partly supported by National Natural Science Foundation of China under Grant No. 61402001.

References

- [1] L. Sweeney, "K-Anonymity: A Model for Protecting Privacy", *International Journal of Uncertainty, Fuzziness and Knowledge-based Systems*, vol. 10, pp. 557-570, May 2002.
- [2] L. Sweeney, "Achieving K-Anonymity Privacy Protection using Generalization and Suppression", *International Journal of Uncertainty, Fuzziness and Knowledge-based Systems*, vol. 10, pp. 571-588, May 2002.
- [3] A. Machanavajjhala and J. Gehrke, Eds., "L-Diversity: Privacy beyond K-Anonymity", in *Proceedings of the 22nd International Conference on Data Engineering*, 2006, pp. 24-35.
- [4] N.H. Li and T.C. Li, "T-Closeness: Privacy Beyond K-Anonymity and L-Diversity", in *Proceedings of the 23rd IEEE International Conference on Data Engineering*, 2007, pp. 106-115.
- [5] X.K. Xiao and Y.F. Tao, "M-Invariance: Towards Privacy Preserving Re-publication Of Dynamic Datasets", in *Proceedings of the SIGMOD*, 2007, pp. 689-700.
- [6] A. Campan and T. Truta, "Data and Structural K-Anonymity in Social Networks", in *Proceedings of the PinKDD*, LNCS 5456, 2009, pp. 33-54.
- [7] J. Cheng and A.W. Fu, Eds., "K-Isomorphism: Privacy Preserving Network Publication against Structural Attacks", in *Proceedings of the international conference on Management of data*, 2010, pp.459-470.
- [8] G. Cormode and D. Srivastava, Eds., "Anonymizing Bipartite Graph Data using Safe Groupings", *The VLDB Journal*, vol. 19, pp.115-139, February 2010.
- [9] K. Liu and E. Terzi, "Towards Identity Anonymization on Graphs", in *Proceedings of the SIGMOD*, 2008, pp.93-106.
- [10] C.H. Tai and P.S. Yu, Eds., "Privacy-Preserving Social Network Publication against Friendship Attacks", in *Proceedings of the SIGKDD*, 2011, pp.1262-1270.
- [11] W.T. Wu and Y.H. Xiao, Eds., "K-Symmetry Model for Identity Anonymization in Social Networks", in *Proceedings of the EDBT*, 2010, pp.111-122.
- [12] L. Zhang and W. Zhang, "Edge Anonymity in Social Network Graphs", in *Proceedings of the CSE*, 2009, pp.1-8.
- [13] L. Zou and L. Chen, Eds., "K-Automorphism: A General Framework for Privacy Preserving Network Publication", in *Proceedings of the PVLDB*, 2009, pp.946-957.
- [14] L. Liu and J. Wang, Eds., "Privacy Preserving in Social Networks against Sensitive Edge Disclosure", in *Proceedings of the SIAM International Conference on Data Mining*, 2009, pp.156-164.
- [15] Y. Li and H. Shen, "Anonymizing Graphs against Weight-Based Attacks", in *Proceedings of the ICDM Workshops*, 2010, pp.491-498.
- [16] S. Bhagat and G. Cormode, Eds., "Class-Based Graph Anonymization for Social Network Data", in *Proceedings of the PVLDB*, 2009, pp.766-777.
- [17] B. Zhou and J. Pei, "Preserving Privacy in Social Networks against Neighborhood Attacks", in *Proceedings of the ICDE*, 2008, pp.506-515.
- [18] M. Yuan and L. Chen, Eds., "Personalized Privacy Protection in Social Networks", in *Proceedings of the PVLDB*, 2010, pp.141-150.
- [19] D. Chakrabarti and Y. Zhan, Eds., "R-Mat: A Recursive Model for Graph Mining", in *Proceedings of the SDM*, 2004, pp.1-5.

Authors	
	<p>Pingshui Wang, born in May, 1972, Suzhou city, P.R. China</p> <p>Current position, grades: Associate Professor of Anhui University of Finance & Economics. University studies: B.Sc. of Computer Science & Technology from Anhui Normal University in China. M.Sc. from Hefei University of Technology in China. D.Sc from Nanjing University of Aeronautics and Astronautics in China. Scientific interest: Information Security. Publications: more than 20 papers published in various journals. Experience: teaching experience of 19 years.</p>
	<p>Xuedong Zhang, born in January, 1980, Benbu city, P.R. China</p> <p>Current position, grades: Associate Professor of Anhui University of Finance & Economics. University studies: B.Sc. of Computer Science & Technology from Anhui Normal University in China. M.Sc. from Suzhou University in China. Scientific interest: Information Security. Publications: more than 10 papers published in various journals. Experience: teaching experience of 12 years.</p>
	<p>Pei Huang, born in July, 1992, Benbu city, P.R. China</p> <p>Current position, grades: Undergraduate of Jilin University in China. University studies: Undergraduate of Jilin University in China. Scientific interest: Information Security. Publications: 3 papers published in various journals.</p>