2010 International Conference on Computational and Information Sciences

Anonymizing Social Network using Bipartite Graph

Lihui Lan
Computer Science School
JiangSu University
ZhenJiang, JiangSu, China
Computer Science School
Jilin Normal University
Siping, Jilin, China
e-mail:lanlihuicaoyue@163.com

Shiguang Ju Hua Jin
Computer Science School
JiangSu University
ZhenJiang, JiangSu, China
e-mail:{Jushig, JinHua}@ujs.edu.cn

***Abstract*—Social networks applications have become popular for sharing information. Social networks data usually contain users' private information. So privacy preservation technologies should be exercised to protect social networks against various privacy leakages and attacks. In this paper, we give an approach for anonymizing social networks which can be represented as bipartite graphs. We propose automorphism publication to protect against multiple structural attacks and develop a BKM algorithm. We perform experiments on bipartite graph data to study the utility and information loss measure.**

Keywords- anonymizing publication; automorphism; bipartite graph; social networks

I. INTRODUCTION

Privacy-preserving data publication has received increasing interest in database community. With the increasing popularity of social network applications, such as facebook and myspace, analysis of these data has also started to attract attention. As a consequence, the amount of social networks data has grown rapidly. It offers rich opportunities for data mining and analysis. Social networks data usually contain users' private information. So privacy preservation technologies should be exercised to protect social networks against various privacy leakages and attacks.

The study of techniques to allow safe anonymization of sensitive data has been ongoing for many years. For example, *k*-anonymity [1] and its variants [2,3,4] are data perturbation techniques designed for tabular micro-data. Some researchers have already studied problems in privacy preserving social networks. Hay et al. [5] and Zhou et al.[6] presented a framework to add and delete some un-weighted edges in social network to prevent neighborhood attackers. Zheleva et al.[7] proposed a model in which nodes are not labeled but edges are labeled which are sensitive and should be hidden. Campan and Truta[8] proposed building "clusters" of nodes, and revealing only the number of edges within a group and between pairs of groups. Cormode et al. [9] studied anonymization of the subclass of bipartite graphs which link two different types of entity, and proposed a permutation-based approach. Lei Zou et al.[10] proposed *k*-automorphism to protect against multiple structural attacks.

In this paper, we give an approach for anonymizing social networks which can be represented as bipartite graphs. We present an algorithm called BKM for anonymizing bipartite graphs, similar to the one described in [10]. We focus on an adversary whose goal is to re-identify a known individual in the anonymized social networks. We show that the algorithm performs well in terms of protection it provides. Table I summarizes the notation used in this paper.

TABLE I. NOTATION

Symbol	Definition
G	an initial social network graph
G_b	a bipartite graph of G
G_{na}	a naive anonymiaztion bipartite graph of G_b
$G^{\#}$	a bipartite graph automorphism publication of G_{na}

II. SOCIAL NETWORK PRIVACY MODEL

A. Social Network Graph

We consider social network graph which describes entities and relationships between entities. A network or graph G is a set of n nodes connected by a set of m edges. The network considered here is binary, symmetric, and without self-loops. In general, we can have different types of nodes and different types of edges in G. For the purposes of this paper, we focus on the case where there are a single node type and multiple edge types.

More formally, we consider a database describing a multi-graph $G = (V, E^1, \ldots, E^k, E^s)$, composed of a set of nodes V and sets of edges $E^1, \ldots, E^k, E^s$. Each node v_i represents an entity and the edge $e^1_{i,j}$ represents a relationship of type E^1 between two nodes v_i and v_j. The graph G shows that there are rich interactions in a real social network.

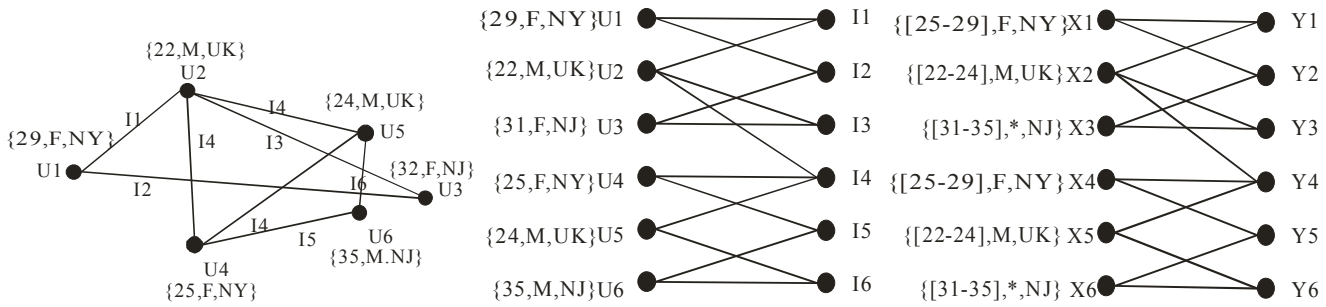
A rich interaction graph G encodes a variety of interactions between a set of entities V. V can represent the members of the social network. The interactions between them can be, for instance, that an email or IM was sent between a pair, a game was played among four players, or a large group declared their support for a political candidate.

B. Bipartite Graph

We present in this paper an anonymization approach for social network data that consists of nodes and relationships. We choose to represent such rich interaction graphs as bipartite

This work was supported in part by National Natural Science Foundation of China under Grant 60773049.

978-0-7695-4270-6/10 $26.00 © 2010 IEEE
DOI 10.1109/ICCIS.2010.245



(a)Original social network graph (b)Bipartite graph model (c)Naive anonymization bigraph

Figure1. Example of social networks

graphs. We divide nodes into two classes V and I. Each node in V represents an individual entity and is described by identifier, quasi-identifier, and sensitive attributes. Each node in I corresponds to an interaction between a subset of entities from V: an edge ($v \in V$, $i \in I$) indicates that the entity represented by node v participates in interaction i. The edge between v and i is unlabeled.

The bipartite graph (bigraph for short) $G_b=(V,I,E)$ consists of $n=|V|$ nodes of one type, $m=|I|$ nodes of a second type, and a set of $|E|$ edges $E \subseteq V \times I$. We shall work with an example of friend network. Fig.1 (b) shows the bigraph representation of the original social network in Fig.1 (a).

C. Anonymization Bigraph

Our objective is to publish an anonymized version of the graph G_b, which still allows a broad class of queries to be answered accurately, but which maintains privacy of the associations. The goal of anonymization is to prevent rediscovering of this sensitive information with a high confidence. We begin by studying naive anonymization bigraph, in which the nodes of G_b are renamed and the structure of the bigraph is unmodified.

Definition 1 (*Bigraph naive anonymization*): The naive anonymization of a bigraph $G_b=(V,I, E)$ is an isomorphic graph, $G_{na}=(V_{na},I_{na},E_{na})$, defined by random bijection $f_v: V \rightarrow V_{na}$ and $f_i: I \rightarrow I_{na}$. The edges of G_{na} are $E_{na} = \{(f_v(x), f_w(y))|(x,y) \in E\}$.

Fig.1(c) shows a naive bigraph anonymization of Fig.1 (b).The adversary does not have direct access to the original graph G, which is hidden. But the adversary may have access to external information about the entities in the graph and their relationships. Faced with the naive anonymized graph, the adversary would like to associate an entity known to be present in G with its representative node in G_{na}.

III. Preserving Privacy Through Automorphism

A. Automorphism Publication

Automorphism publication induces a partitioning on G_{na} into sub-graphs. These partitions are isomorphic graphs and have identical structural properties. It follows that an adversary, even with exhaustive knowledge of a target node's structural position, cannot isolate an individual beyond the set of entities to which it is automorphically equivalent. We formalize structural query describing the external information available to an adversary.

Definition 2 (Structural query): Given an anonymization bigraph $G^{\#}=(V^{\#},I^{\#},E^{\#})$, a query Q refers to any information that an attacker can use to extract private information from $G^{\#}$. For an entity $x \in V$, called the target, its candidate set contains the nodes of $G^{\#}$ that could feasibly correspond to x. The result of Q is a candidate set that a set of vertices $V' \subseteq V^{\#}$ and each $v_i \in V'$ is called a match vertex to x, denoted by *CandSet(x)*.

Our goal is that there are at least k candidate nodes for any node x in the original data through automorphism publication.

Definition 3 (Bigraph automorphism): An automorphism of a bigraph $G_b = (V,I,E)$ is an automorphic function f of the vertex set V and I, such that for any edge $e= (v,i)$ ($v \in V, i \in I$), $f(e) = (f(v),f(i))$ is also an edge in G_b. *If there exist k automorphisms in G_b, it means that there exists k-1 different automorphic functions.*

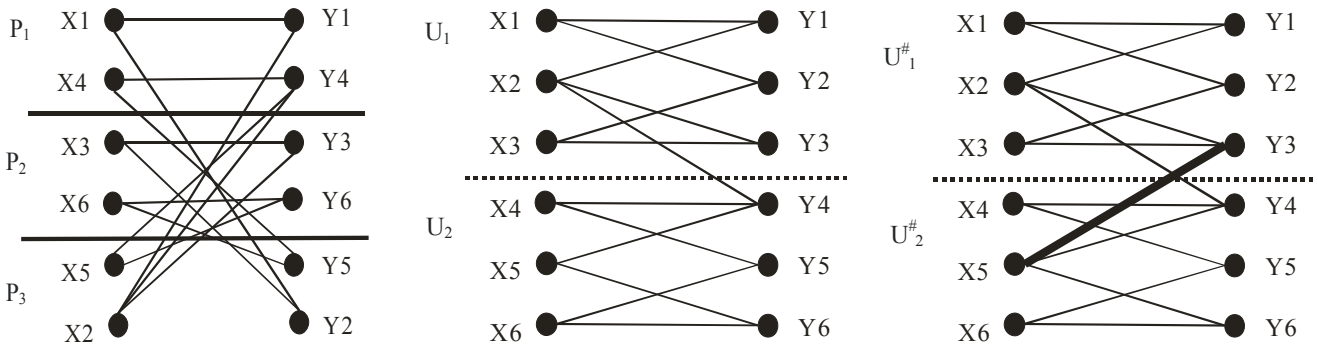
Definition 4(Sub-Graph Isomorphism[10]): Given two graphs Q and G, if there exists at least one sub-graph X in graph G such that Q is isomorphic to X under the bijective function f, graph Q is sub-graph isomorphic to graph G. We call X a sub-graph match of Q in G. The vertex $f(v)$ in G is called the match vertex with regard to vertex v in Q.

Definition 5(Automorphism publication): Given a bigraph G_b, $G^{\#}$ is automorphism publication to G_b, *if and only if* G_b *is* sub-graph isomorphism to $G^{\#}$ and $G^{\#}$ is automorphism bigraph.

B. BKM algorithm

Given an initial bigraph modeled as a graph $G_b =(V,I,E)$, nodes from V are described by quasi-identifier and edges from E are undirected and unlabeled. The algorithm described in this section, called the BKM (*Bigraph K-Automorphism Match*), finds an anonymized bigraph $G^{\#}$ that satisfies *k-automorphism* match principle. Obviously, if a released bigraph $G^{\#}$ satisfies *k-automorphism*, given any structural query Q, no adversary can identify the target with a probability higher than $1/k$.

We illustrate the main idea of BKM algorithm using Fig. 2. First, the algorithm establishes a partitioning of all nodes from V into clusters. Let us return to the earlier example, suppose k



(a)Partition bigraph into blocks (b) Divided bigraph into groups (c) Bigraph publication

Figure2. Example of bigraph automorphism publication

Algorithm: BKM algorithm
Input: An original social network graph G and the parameter k.
Return: The anonymized bigraph $G^{\#}$, which is a *k-automorphism* .
1. Construct bigraph model G_b from G.
2. Generate a naive anonymization bigraph G_{na} from G_b.
3. Generalize nodes V attributes' values according to k to obtain n blocks P_j, ($j = 1,..., n$).
4. Select one node v_j from P_j to form into m groups $U_i=\{V_i,I_i,E_i\}$, ($i = 1, ...,m$), $V_i=\{v_1,...v_n...v_n\}$,$v_1 \in P_1,..., v_i \in P_i$, where each V_i has at least n nodes .
5. Perform graph isomorphic on all groups U_i to obtain $U^{\#}_i$.
6. Replace each group U_i by $U^{\#}_i$.
7. For all crossing edges, perform edge-copy to obtain anonymized bigraph $G^{\#}$.
8. Return $G^{\#}$.

Figure3. BKM algorithm

is set to *2* and guarantee that the released bigraph $G^{\#}$ satisfies *2-automorphism*. The generalization of the quasi-identifier attributes is one of the techniques widely used for micro-data[1,2,3,4]. We reuse this technique for the generalization of nodes V attributes' values. We partition the original network G_{na} into *3* blocks, P_1 ,P_2 and P_3, as shown in Fig. 2(a). Second, we select one node v_j from P_j ($j=1,2,3$) to form into *2* groups, U_1 and U_2, as shown in Fig. 2(b). Third, we perform graph isomorphism on U_i ($i=1,2$) to obtain two groups $U^{\#}_1$ and $U^{\#}_2$. Last, for all crossing edges, perform edge-copy to obtain an anonymized bigraph $G^{\#}$(add an edge (X5,Y3)) in Fig.(c) .The BKM algorithm shows in Fig.3.

In theory aspect, the utility of BKM algorithm depends on how symmetrical the original social network graphs are. If there are many automorphism partitions [11] with no fewer than k vertices in original graph, we will introduce few noisy edges. The utility of BKM will be good in this case; otherwise, the utility will degrade. Many real networks are known to have high symmetry property [11, 12].

C. Information Loss

In this paper, we use two information loss measures. One quantifies how much descriptive data detail is lost through quasi-identifier attributes generalization–we call this metric the generalization information loss measure. The second measure quantifies how much structural detail is lost through isomorphic graph construction and it is called structural information loss.

The generalization of quasi-identifier attributes reduces the quality of the data. We use the generalization information loss measure as introduced and described in [8]. Information loss quantifies the probability of error when trying to reconstruct the structure of the initial social network from its publication version. We introduce a measure to quantify the structural information which is lost when anonymizing a bigraph through constructing isomorphic graph. We consider the number of added edges in anonymous bigraph as structural information loss measures. Given an original bigraph G_b and its anonymized version $G^{\#}$, the anonymization cost is defined as $Cost(G_b) = |E(G^{\#})| - |E(G_b)|$, where $E(G_b)$ is the set of edges in G_b. The normalized generalization information loss [8] and structure information loss definitions show in (1) and (2).

$$NGIL(G_b) = \frac{GIL(G_b)}{n(s+t)} \quad (1)$$

$$NSIL(G_b) = \frac{Cost(G_b)}{|E(G_b)|} \quad (2)$$

D. Experimental Results

In this section, we evaluate the utility of the anonymized data through experiments on the synthetic data. The algorithms were implemented in Java; tests were executed on a CPU machine with 2.0GHz and 1GB of RAM, running Windows XP Professional. We used the random data generator to generate the synthetic data sets. In original social network graph, there are $|V|$=186 and $|E|$=584. In the experiment, we considered a set of four quasi-identifier attributes: age, race, sex, and native country. Fig. 4 presents the normalized structure information loss and normalized generalization information loss with the different values of k by applying the BKM algorithm.

Experiments show that our method has higher data utility for various types of attacks. Existing methods assume a single type of attack except [10]. However, reference [10] represents a social network as a standard graph. In social network publication, it is important for the anonymization to mask the associations between entities and their interactions. Bipartite

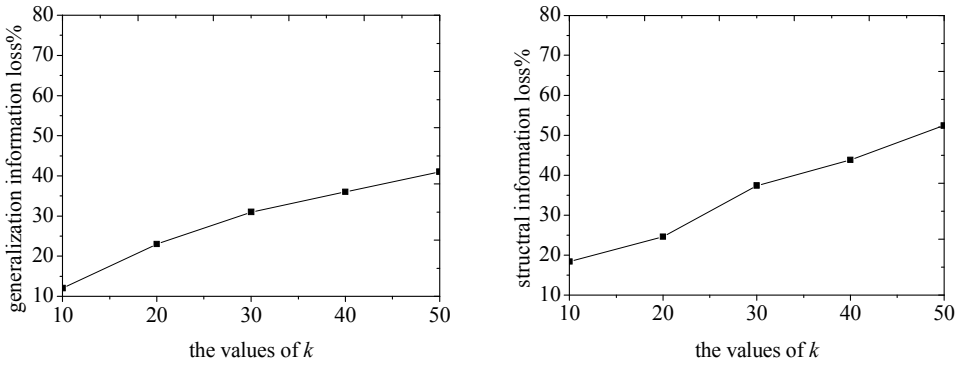


Figure4. Experimental results

graph can better than standard graph in hiding relationships between entities. Our method can guarantee privacy under any structural attack. Therefore, it provides much stronger privacy protection than the others.

IV. CONCLUSION

The availability of digital technologies and internet development has promoted a proliferation of social networks. Due to the public awareness of privacy protection, the sharing potential of certain social networks may be seriously hampered by the need for a balance between the protection of sensitive content and public availability of data utility. In this paper, we studied an anonymization approach for social networks data. Our focus has been on data that can be represented as a bipartite graph linking two types of entity. We developed an algorithm called BKM that anonymizes a bigraph through automorphism. This algorithm can be user-balanced towards preserving more the structural information of the bigraph or the nodes' attribute values. We introduced a measure to quantify information loss in bigraph publication.

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, April 2002.

[2] A.Machanavajjhala,J.Gehrke, D.Kifer, and M.Venkitasubramaniam , "*l*-diversity: Privacy beyond k-diversity," in the 22nd International Conference on Data Engineering. New York:ACM,2006,pp. 24-35.

[3] N.Li,T.Li,and S.Venkatasubramanian, "*t*-closeness:Privacy beyond *k*-anonymity and *l*-diversity, " in the 23rd International Conference on Data Engineering. Istanbul:IEEE ,2007,pp.106-115.

[4] HAN Jian-min,CEN Ting-ting,YU Hui-qun, "Research in Micro-aggregation Algorithms for k-Anonymization, " Chinese Acta Electronica Sinica, vol.36,pp. 2021-2029,November 2008.

[5] M.Hay, G.Miklau, D.Jensen, P.Weis, and S.Srivastava, "Anonymizing social networks, " University of Massachusetts Amherst,Tech. Rep. 07-19, 2007.

[6] B.Zhou and J.Pei, "Preserving privacy in social networks against neighborhood attacks, " in the 24th International Conference on Data Engineering. Washington DC:IEEE ,2008, pp.506-515.

[7] E.Zheleva and L.Getoor, "Preserving the privacy of sensitive relationships in graph data, " in the First ACM SIGKDD international Workshop on Privacy,Security,and Trusting KDD. Berlin :Springer, 2007, pp.153-171.

[8] A.Campan and T.M.Truta, "A clustering approach for data and structural anonymity in social networks, " in the 2nd ACM SIGKDD International workshop on Privacy, Security, and Trust in KDD. Lasvegas:ACM,2008,pp.1-10.

[9] G.Cormode,D.Srivastava,T.Yu,and Q.Zhang, "Anonymizing bipartite graph data using safe groupings, " in the 34th International Conference on Very Large Databases. Auckland : VLDB Endowment,2008,pp. 833-844.

[10] ZOU Lei, CHEN Lei, and ÖZSU M T, "K-Automorphism: General Framework for Privacy reserving Network Publication, " Proceedings of the VLDB Endowment,vol.2,pp. 946-957, January 2009.

[11] J. Lauri and R. Scapellato, "Topics in Graph Automorphisms and Reconstruction, " Cambirdge University Press,2003.

[12] Y. Xiao, M. Xiong, W. Wang, and H. Wang, "Emergence of symmetry in complex networks, "Physical Review E, vol.77,pp.1-10, June 2008.