A novel clustering-based anonymization approach for graph to achieve Privacy Preservation in Social Network

Huowen Jiang[1, a]

[1]School of Mathematics & Computer Science, Jiangxi Science & Technology Normal University, Nanchang, 330038, China

[a]Jhw_604@163.com

Keywords: Clustering anonymity, k -anonymity graph , privacy preservation, social network

Abstract. Serious privacy concern rises with the prosperity of social network applications. To prevent the privacy of vertices or edges associated with entities in a social network from getting re-identified through background information or queries,a novel clustering-based approach is proposed to anonymize vertices and edges. Concepts of vertex similarity matrix and the distance between a vertex and a cluster are defined, based on which a k -anonymized graph approach is presented. The effectiveness of the approach is verified Through experiments that compare the performance of our method with that of SASN, an existing anonymization algorithm .

1 Introduction

Along with the rapid development of Mobile Internet and Internet of things, social networks based on network technology have attracted a wide range of people's attention. Because of its prosperity and wide usage, more and more scholars and engineers are focusing their research and application developing on social network. Analysis on social network has become a hot spot in sociology, informatics, geography, economics and many other research fields[1] . However, data of social networks are released in large number to meet the requirement of science research and social sharing, which , meanwhile, causes a serious problem of private information misuse. Privacy preservation on social network applications has attracted huge concerns of people from industry leaders to ordinary users. For social network service providers, one of the key points is how to make social network serve the public without making users' privacy known. For one thing, releasing raw data directly will disclose privacy of individuals and groups, which means data should be processed before releasing; for another, the processed data should be still practically useful for third-party institutes to analyze and work on. Therefore, the degree of privacy preservation and the usability of data should be well balanced so that data can go to public safely and serve the society[2].

In recent years, researches concentrating on privacy preservation in social network are emerging in the ascendant. k -anonymity has been widely used to achieve privacy preservation for relational data since its proposal by Samarati and Sweeney[3] . Approaches based on K-anonymity were presented to protect privacy for network graph data by realizing vertex k -anonymization or subgraph k -anonymization [4,5]. Focusing on achieving vertex k -anonymization privacy preservation, we put forward a clustering approach based on vertex similarity. To be more exact, in our approach, a social network graph is partitioned into several super vertices, each of which contains at least k vertices that cannot be distinguished from each other, making the probability of the subject of one vertex being re-identified less than $1/k$.

2 The social network graph and its privacy prservation

A social network, also called a social relational network, refers to a relatively stable social relations system made up of a set of social entities (individuals and organizations) and a set of dyadic ties between these entities. A social network is organized as a map of relations, which means it makes sense to model and analyze a social network using a graph where vertices represent the entities in the network while edges represent the ties between these entities.

Def.1(social network graph): A social network graph refers to a graph comprising a finite number of vertices and edges that characterize entities and ties between these entities in a social network. A social network can be represented by $G = G(V, E)$, where V is the set of vertices and E is the set of edges in graph G.

For instance, G_1 in Fig.1 shows a simple undirected graph without tagged information, where V ={A,B,C,D,E,F} and E ={<A,B>,<A,C>,<C,D>,<D,E>,<D,F>,<E,F>}. In addition to graph, a social network can be also characterized in matrix form, called adjacency matrix.

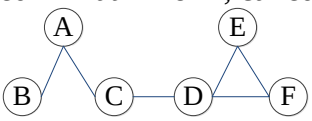

Fig. 1 G_1:A simple undirected graph

Def.2(adjacency matrix): Supposing $G = G(V, E)$ is a simple graph with n vertices numbered from 1 to n, $A(G) = (a_{ij})_{n \times n}$ is said to be the adjacency matrix of G if $A(G)$ is a square matrix of order n, in which a_{ij} =1 if there is an edge between vertex i and vertex j and a_{ij} =0 if there is not.

For instance, the social network shown in Fig.1 can be also represented by the adjacency matrix in Fig.2, which is symmetric because the social network is undirected.

$$\begin{pmatrix} 0 & 1 & 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 & 1 & 1 \\ 0 & 0 & 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 1 & 1 & 0 \end{pmatrix}$$

Fig.2 The adjacency matrix of graph G_1

Nowadays, social networks are widely used in the Internet such as QQ, Wechat, Blog, Microblog, Facebook, Myspace, Twitter and etc. The data held by a social network can be categorized into two types: one is attribute data which carry information of entities that can be used to identify the subject of these entities, such as name, alias, id number and etc.; the other is relational data that record ties and interactions between entities. Social network data are often large capacity, high-dimensional, non-linear. Since we model a social network with a graph, existence and attribute values of the vertices and edges, topology and other properties of a graph are endowed practical meanings and can be used to access private information by attackers. To prevent directly exposing users' identity, data carried by vertices in a social network is often anonymized. A graph is said to be a simple anonymized graph if just the explicit identifiers of the vertices are hidden. For instance, the graph shown in Fig.3 is a simple graph of the one in Fig.1. And yet if attackers manage to get the knowledge like the degree or neighborhood information of certain vertex by background information or other sources, the probability that attackers can re-identify the subject of the entity represented by the vertex is high. For example, suppose the graph in Fig. 1 characterizes a friend circle and if an attacker knows Bob has at least three friends, then it is not hard to infer vertex 4 represents Bob. To address this issue, we anonymize $G^{'}$ one more time, partitioning vertices into clusters and constructing k -anonymized graph G^*.

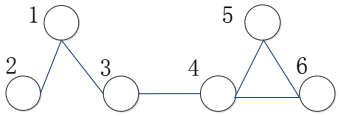

Fig.3 $G_1^{'}$:A simple anonymized graph of G_1

Fig.4 G_1^*:A k -Anonymized graph

Def.3(k -anonymized graph): For a simple anonymized graph $G^{'}$, graph $G^* = G^*(V^*, E^*)$ is said to be its k-anonymized graph if all vertices in $G^{'}$ are partitioned into clusters according to their similarity and each cluster is represented by a vertex in G^*, called a hyper-vertex and written （$\left|V^*_{\mathrm{i}}\right|$,

$\left|E^*_{\mathrm{i}}\right|$), where V^* is a set composed of finite hyper-vertices, $\left|V^*_{\mathrm{i}}\right|$ and $\left|E^*_{\mathrm{i}}\right|$ represents the number of hyper-vertices and the number of edges respectively. For instance, the k-anonymized graph of the simple graph in Fig.1 is shown in Fig.4.

3 The similarity and distance between vertices

k-anonymized graph can prevent privay disclosure of some vertices or edges associated with individual entities in a social network.so the processing of k-anonymization for social network graph is an effective method to protect private information of individuals in social network. Clustering is one of the k-anonymity means, and the measurment of similarity and distance between vertices is an important basis for clustering. Referring to definition and calculation of vertex similarity and distance in Ref.[7], we give the formulas of calculating the similarity degree between vertices and distance between vertex and vertex or cluster as follows:

The formula 1: $E = (I + A)^m$ ····· (1)

where E represents the vertex similarity matrix, I represents the identity matrix of order n , A represents the adjacency matrix, and m represents a performance parameter. It was found that the result is more precise when m =3 or m =4, we suppose m =3.

The formula 2: $dist(v_i, v_j) = \dfrac{\sum_{\mathrm{k}=1}^{n}(\left|e_{i,k} - e_{j,k}\right|)}{n}$ ····· (2)

where $dist(v_i, v_j)$ represents the distance between vertices v_i and v_j , $e_{i,k}$ represents the value of the i-th row, j-th column element of the similarity matrix A, n is the number of graph vertices.

The formula 3: $dist(v_i, C_j) = \dfrac{\sum_{\forall v_t \in C_j} dist(v_i, v_t)}{\left|C_j\right|}$ ····· (3)

where $dist(v_i, C_j)$ represents the distance between vertices v_i and cluster C_j, $\left|C_j\right|$ represents the number of vertices in cluster C_j. Obviously, $\left|C_j\right| \geq k$.

4 The clustering-based anonymization approach for graph

Based on the definition of the distance between vertices and the distance between vertex and cluster introduced in section 3, this section put forward a k-anonymity algorithm based on clustering (KAAC) so as to protect privacy in social network. The basic idea of the algorithm is as follow: The graph of n vertices is divided into a number of clusters using clustering ideas, meanwhile make sure that each cluster contains at least k vertices. The subgraph,which is composed of all vertices and edges within a cluster, can be replaced by a super vertex. As long as existing a connecting edge between the any two original subgraphs, the corresponding two super vertices have an edge connected. The method that divides vertices into a cluster is as follow: select any one non-clustered vertex as a starting item to re-form a cluster; select a vertex from non-clustered vertices according to the principle of minimizing the distance between a vertex and the cluster, and then adds it to the cluster. If there are multiple vertices that have the minimum distance to the cluster at the same time, then select all of them until the number of vertices of the cluster is greater than or equal to k . Detailed description of the KAAC algorithm is given as follow:

Algorithm 1: A clustering-based anonymization Algorithm for social network graph.

Inputs: a social network graph G , the adjacency matrix of graph G , anonymity parameter k .

Outputs: k-anonymized graph G^* .

Steps:

1. select any one vertex from G to form a new cluster C_t ;

2. calculating the distance of all non-clustered vertices to C_t, adding a vertex that has the smallest value; or mutliple vertices in case of having the same smallest distance , to C_t;
3. repeating step 2 until the number of vertices of cluster C_t is greater than or equal to k;
4. repeating step 1~step 3 until the number of vertices of G is less than k;
5. selecting any one vertex from the remaining non-clustered vertices, calculating respectivly the distance to every clusters, then adding the vertex to the cluster that has the minimum distance value;
6. repeating step 5 until there is no non-clustered vertices, and then all vertices are clustered;
7. calculating the number of vertices and edges in every clusters, replacing them with a corresponding super vertex seperatly, and finally output the k-anonymized graph G^*.

5. Experiment and its analysis

To prove our theory, in this chapter, we conduct experiments to analyze the practical efficiency of KAAC and compare it with SASN proposed in the Ref.[7].The design of data used in our expermients refer to Ref.[7]. The experiment is carried out in such hardware environment: Intel Pentium double-core E2140 @1.60GHz CPU,2GB (DDR) memory. The algorithm is implemented on Microsoft Visual C++ 7.0. Two experiment is conducted to respectivly reveal the runtime how to be changed with increasing the value of k, assuming m =3 and 4 separately in formula 1. In view of that picking different initial vertex to initialize a new cluster may lead to different result, each experiment runs 10 times and the final result of each experiment is the average of the 10 outcomes.The final rusults of two expermients are shown in fig.5 and fig.6 respectivly.

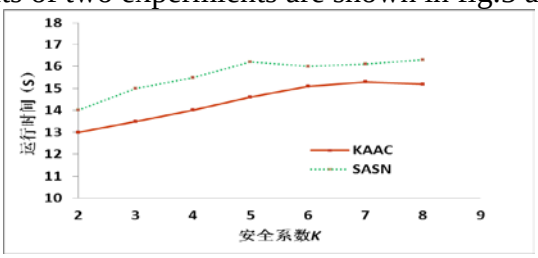

Fig.5 m=3， runtime changed with k

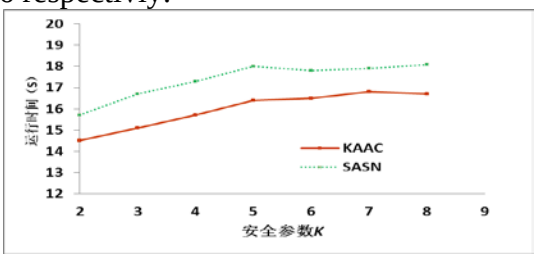

Fig.6 $m=4$， runtime changed with k

The results above reveal:①With the increase of k value,the runtime of the algorithm is slightly increasing,and the runtime of the KAAC is shorter than the SASN.This can be explaind,due to increasing the value of k ,the much more vertices are included in the cluster when establishing one cluster, the runtime of establishimg a cluster will be increased;on the other side, the number of the clusters is reduced because of the total of vertices is a constant,so the general runtime only have slight variations.②Comparing the rusults in fig4. and fig.5,the runtime of two algorithm in fig5. have more increased slightly than it in fig4.This is because two algorithms have much more computation when coming to the simirality martiex of vertices when $m=4$. Obvisuoly the experiment show that the algorithm of KAAC is feasible.

6 Conclusions

k-anonymizing the corresponding social network graph is an important way to achieve privacy preservation for a social network. Private information of subjects carried by vertices and edges in a social network graph may be accessed by attackers who have certain background information about the social network. To address this issue, we propose KAAC, a novel clustering-based anonymization approach, which transfers the simple anonymized graph of a social network to a k-anonymized graph, effectively preventing the privacy disclosure mentioned above. It is undeniable that our method hides connectivity of some vertices, which changes the structure of a social network graph to some extent and brings up information loss. Therefore, we will focus on how much

information loss is caused by the anonymization and how to balance the information loss and the degree of privacy preservation in our further research.

Acknowledgements

This research work was supported by the Project of Science & Technology Plan of Jiangxi Provincial Education Department under grant No.GJJ13569. the Subject of Teaching Reform in Universities of Jiangxi Provincial Education Department under grant No. JXJG-14-10-9.

References

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

[2] Wu XW. Research on anonymity techniques for privacy-preserving data publishing in social networks[D]. Harbin Engineering University, 2013.

[3] L. Sweeney. *k* -anonymity: a model for protecting privacy. Int'l Journal on Uncertainty, Fuzziness, and Knowledge-Based Systems, 2002, 10(5):557-570.

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

[5] Bhagat S, Cormode G, Krishnamurthy B, Srivastava D. Class-Based graph anonymization for social network data. In: Proc. of the 35th Int'l Conf. on Very Large Databases, 2009:766-777.

[6] Wang Y. Research on privacy preservation of social network[D]. Nanjing University of Posts and telecommunications, 2013.

[7] Xiang KL. Research on privacy preserving in social networking based on graph modification and clustering[D]. Zhejiang University, 2013.