A Personalized $a(d, k)$-Anonymity for Social Network

Xiang-min REN*, De-xun JIANG, Ke-chao WANG and Qi RAN
School of information engineering, Harbin University, Harbin, China
*Corresponding author

Keywords: K-anonymity, Privacy protection, $A(d, k)$-anonymity.

Abstract. Privacy protection of Social Network becomes a more and more serious concern in applications. The common way to protect privacy is to use $k$-anonymity in data publishing of social network. In this paper, we study the theory of $k$-anonymity of social network, use weighted lower triangular matrix to represent the relationship among the nodes of social network, and propose a personalized $a(d, k)$-anonymity mode. The $a(d, k)$-anonymity algorithm experiments prove that it can make anonymous nodes of social network effectively resist $d$-neighborhood attack, and structure attack, at the same time, make the node information availability according to the personalized weight parameter $a$. The algorithm is usable, effective and efficient comparing with traditional $(d, k)$-anonymity algorithm.

Introduction

Social network is playing more and more important role in our society today. More and more people can make friends, share information, get news, play games and so forth on social network such as Facebook, SINA Weibo, Tencent Weibo, and so on. There are a large number of individual data and the enterprise data in social network, but individual data and the enterprise data including their sensitive or private data are possibly leaked at any moments when publishing and sharing the information on social network. In order to guarantee the information security on social network when publishing and sharing data, a lot of scholars research and develop the social network publishing data security technology on different levels. Kumar and Novak proposed Social network structure and evaluation method [1], attack types of privacy information of social network have been fully researched in paper [2], different types of data including sensitive information and graph were proposed in paper [3]. Some anonymity data publishing methods and techniques in relational database can be improved in social networks [12].

$K$-anonymity of social network must satisfy that there are at least $k-1$ nodes with the same structure for the sensitive individual although adversary has background knowledge of social network structure, which make adversary identify individual with probability of no greater than $1/k$. Fig.1 is an example of $2$-anonymity publishing graph of social network, Fig.1(b) satisfies $2$-anonymity publishing graph by replacing name with 1,...,5 and adding an edge, that is to say, although adversary knows sensitive individual (node 5) degree, and he couldn’t identify node 5 with no greater than $1/2$.

![Figure 1. 2-anonymity publishing in a social network.](image)

More and more anonymity models and methods of social network were proposed such as $k$-degree anonymity [4-7], $k$-neighborhood anonymity [8], $k$-isomorphism anonymity [9], etc. $K$-degree anonymity can make all nodes have the same degree, and can protect privacy, $k$-neighborhood anonymity can resist attack of adversary with neighborhood relationship background knowledge, $k$-automorphism can resist attack of adversary with background knowledge of any subgraph including
destination node, \( k \)-symmetry can resist attack of adversary with background knowledge of any sub-graph including destination node, \( k \)-isomorphism anonymity can resist structural attacks of adversary.

In this paper, we propose a \((d, k)\) -anonymity model of social network, it can not only resist structural attacks, but also resist attack of sensitive edges with background knowledge.

**Problem Definition**

**Related Concepts of Anonymity and Graph**

The following is the definitions of several terms mentioned in the paper [10-11].

**Definition 1. \( k \)-degree Anonymity.** A social network graph \( G (V; E) \) is \( k \)-degree anonymity, if each node at least has \( k-1 \) other nodes with the same degree in the same social network graph. \( V \) is numbers of vertices (nodes), \( E \) is numbers of edges between vertices.

\( k \)-degree anonymity can resist the adversary who has background knowledge of node degree. In figure 1, the collection of degree is \( d={2,2,2,2,1} \) in original social network graph (a), and the set is \( d={2,2,2,2} \) in 2-anonymity social network graph (b) by adding one edge between node 2 and 5.

**Definition 2. Graph Isomorphism.** For two graphs \( G_1 = (V_1, E_1) \) and \( G_2 = (V_2, E_2) \) where \( |V_1| = |V_2| \), if there is a bijection \( h \) between \( V_1 \) and \( V_2 \) satisfies: \( \forall (u, v) \in E_1 \), if and only if \( \exists (h(u), h(v)) \in E_2 \), \( G_1 \) and \( G_2 \) are graph isomorphism, denoted as \( G_1 \cong G_2 \). \( V \) is numbers of vertices (nodes), \( E \) is numbers of edges between vertices. For example, when we delete the nodes information of (a) and (b), delete the edge between node 2 and node 5 of (b) in figure 1, (a) and (b) are isomorphic.

**Definition 3 (Subgraph Isomorphism).** For two graphs \( G_1 = (V_1, E_1) \) and \( G_2 = (V_2, E_2) \), if \( \exists G_1_i \subseteq G_1 \) and \( G_2 \) are isomorphic, we say there is subgraph isomorphism from \( G_1 \) to \( G_2 \).

For example, there are subgraph isomorphism from (b) in figure 1 to (a) and (b) in figure 2.

![Figure 2. Example of sub-graph.](image)

**Definition 4 \( d \)-Neighborhood Subgraph.** For a vertex \( v \in V \) in an undirected graph \( G = (V, E) \), \( \forall u \in V \), if the shortest path \( \text{path} (u, v) \) between \( u \) and \( v \) is less than or equal to a positive integer \( d \), then vertex \( v \), all of \( u \), and all the edges along the path \( (u, v) \) is defined as \( d \)-Neighborhood Subgraph of \( v \), denoted as \( \text{Neighbor}_d (v) \).

If \( d=1 \), we say neighborhood subgraph of vertex \( v \) is \( \text{Neighbor}_1 (v) \). For example, (b) is 1-Neighborhood Subgraph of Cindy, (c) is 2-Neighborhood Subgraph of Cindy in Fig. (3).

![Figure 3. Example of d-Neighborhood Sub-graph.](image)

If \( d=1 \), we say neighborhood subgraph of vertex \( v \) is \( \text{Neighbor}_1 (v) \). For example, (b) is 1-Neighborhood Subgraph of Cindy, (c) is 2-Neighborhood Subgraph of Cindy in figure 3.
Sensitive Degree of Friend Relationship

Everyone has a lot of friends, (a) of figure 4 is an example of friend relationship graph, “1” represents simple friend relationship between two nodes, “2” represents good friend relationship between two nodes, “3” represents sweetheart relationship between two nodes, “0” represents no relationship between two nodes. Usually, we divided edge between two nodes into two categories, sensitive edge and non-sensitive edge. But how can you identify good friend relationship or sweetheart friend are sensitive edge? The degree of sensitive is uncertain and different, for example, in friend relationship graph, and we described relationships of simple friend, good friend, sweetheart friend (boyfriend or girlfriend) among the nodes. Some people do not like others to know that he is in the friend relationship graph, some people may ignore it, some people do not like others to know that he has good friends or sweetheart friends, and some people think that it will be ok. Particularly, if sweetheart friend includes gay or lesbian relationship, most of people do not want others to know that he is gay or she is lesbian, so different people have different sensitive degree about friend relationship, and we must meet the needs of personalized privacy protection according to the practical application.

How to solve the problem of sensitive degree difference of friend relationship? We can set up an appropriate threshold of sensitive degree of friend relationship, and we must protect privacy of nodes, when the nodes sensitive degree of friend relationship is greater than the threshold given in advance. The threshold of degree of sensitive friend relationship can be given by expert or fuzzy function. In this paper, we have six relationships in our experiments, sweetheart friend relationship (threshold 4) including boyfriend, girlfriend, gay and lesbian was represented sensitive edge.

Weighted Lower Triangular Matrix

D-Neighborhood Subgraph can be described by weighted lower triangular matrix. We create a matrix including all nodes of friend relationship graph, column vector includes query node, row vector does not include query node, the elements of matrix denote friend relationship (edge) weight among nodes, “0” denotes no relationship, “1” denotes simple friend relationship, “2” represents good friend relationship, “3” represents sweetheart relationship between two nodes, we use \( N_d(v) \) to denote node \( v \) of \( D \)-Neighborhood Subgraph lower triangular matrix. We can easily know the sensitive degree of friend relationship by using weighted lower triangular matrix. For example, 1-Neighborhood Subgraph and 2-Neighborhood Subgraph of node Cindy in figure 4 as follow:

\[
\begin{bmatrix}
A & B & C & D \\
B & 1 \\
C & 1 & 0 \\
D & 2 & 0 & 0 \\
E & 3 & 0 & 0 & 2
\end{bmatrix}
\]
A \((d, k)\)-Anonymity Graph Model

In order to make it impossible for an attacker to infer the real relationship between targeted individuals and corresponding nodes with a probability, \(k\)-anonymity concept in data tables and the new concept of \(a\,(d, k)\)-anonymity are introduced[10].

Definition 5 \(a\,(d, k)\)-anonymity of the Vertex. For an undirected graph \(G = (V, E)\), the graph \(G_p = (V_p, E_p)\) is as its anonymous publishing graph, if a vertex \(v \in V\), there are at least \(k-1\) nodes \(u_1, u_2, \ldots, u_{k-1} \in V_p\) in \(G_p\), which makes \(\text{Neighbor}_v(u) \equiv \text{Neighbor}_v(u_i)\) and \(v \neq u_i\), wherein, \(i = 1, 2, \ldots, k-1\), thus, the vertex \(v\) is \((d,k)\)-anonymity, and the vertex \(v\) is \(a(d,k)\)-anonymity according to personalized parameter \(a\), \(a\) is the weight of relationships(edge weight) of \(d\)-neighborhood of vertex \(v\).

For example, in figure 4, \(a = \{0,1,2,3\}\) of node A(Cindy), and \(a = \{0,1,2,3\}\) of node H(Mary), so node A and node H satisfy \(a(1,2)\)-anonymity.

Definition 6 \(a(d, k)\)-anonymity of the Graph. For an undirected graph \(G = (V, E)\), the graph \(G_p = (V_p, E_p)\) is as its anonymous publishing graph. If any vertex \(v \in V\) is \((d,k)\)-anonymity, thus, the graph \(G_p\) is \((d,k)\)-anonymity, if any vertex \(v \in V\) is \(a(d,k)\)-anonymity, thus, the graph \(G_p\) is \(a(d,k)\)-anonymity.

Subgraph Isomorphism and \(a\,(d, k)\)-anonymity

Based on the Definition 5, there are at least \(k\)-1 different nodes that satisfy \(d\)-neighborhood isomorphism sub-graph in the publishing graph for the \(a\,(d,k)\)-anonymity of the Vertex. If \(d\)-neighborhood subgraph coding is represented with lower triangular matrix, the isomorphism of \(d\)-neighborhood subgraph is considered as the xor matching between matrices. Hence, the process of neighborhood anonymizing between two vertices can be expressed as the following:

```
Algorithm \(a(d,k)\)-anonymity algorithm for social network
Inputs: Weighted anonymous graph \(G=(V, E)\), neighborhood parameter \(d\), anonymous parameter \(k\), weighted parameter \(w_{ij}\).
Outputs: Anonymous graph \(GP=\{g_1, g_2, \ldots, g_e\}\)
Steps:
1. Initialize \(G\), \(GP\) and nodes candidate set \(candSet\).
2. The shortest paths of each node in \(G\) are calculated, and \(G\) is pruned based on \(w\) to get \(G'\).
3. Based on \(d\), the neighborhood sub-graphs of all the nodes are got as \(N_d(V)\).
4. The nodes are sorted based on the sum of 1\(^st\) row in \(N_d(V)\) and saved in \(sortArray\).
5. While \((sortArray \neq \emptyset)\) do
6. \(\text{curVertex} = \text{the first node in } sortArray\).
7. Remove first node in \(sortArray\).
8. \(\text{curVertex} = \text{curVertex} \cup \{\text{candSet}\}\).
9. if \((sortArray\).size \geq 2k-1\) then
10. for \(i=1\) to \(2k-1\) do
11. Compute the value of \(\text{Cost(curVertex, sortArray[i]))}\)
12. end for
13. \(tempSet=\{k-1\ \text{Nodes with the least Cost}\}\)
14. else
15. \(tempSet=\{\text{the rest of nodes}\}\)
16. end if
17. \(candSet = candSet \cup tempSet\)
```
If $N_i(v)\oplus N_j(v) = 0$, then $\text{Neighbor}_i(v) = \text{Neighbor}_j(v)$, otherwise, modify the corresponding matrix element relationship weight value from 0 to $a$, which means to add one edge or add one relationship in the neighborhood subgraph.

For example, in Figure 4(b), we know:

$$N_i(F) = \begin{bmatrix} 3 \\ 2 \\ 1 & 0 \end{bmatrix}$$

(3)

$$N_i(H) = \begin{bmatrix} 3 \\ 2 \\ 1 & 0 \end{bmatrix}$$

(4)

$$N_i(F) = \begin{bmatrix} 3 \\ 2 \\ 1 & 0 \end{bmatrix}$$

(5)

In the formulas above, $N_i(H) \oplus N_i(F) = 0$, thus, the node F and node H are 1-neighborhood subgraph isomorphism. But in this situation, then $N_i(H) \oplus N_i(F) = 1$, which means that the node F and node H are not 1-neighborhood subgraph isomorphism with weighted edges, therefore, modifying the weight between node B and node G from “2” to “1” can achieve 1-neighborhood subgraph isomorphism, node F and node H satisfy $a(1,2)$-anonymity.

D-neighborhood and Structure Attack of Graph

Suppose $G_p$ is an anonymous publishing graph of a social network graph $G$, if the attacker knows the $d$-neighborhood subgraph information of the vertex $v$ corresponding to the target individual A in graph $G$, and the real information corresponding to target individual A in graph $G_p$, that is called $d$-neighborhood attack of graph, or structure attack of graph. For example, in figure 4, if the attacker know Cindy have four relationships (weighted edges), he can find that node A is Cindy easily. Obviously, $a (d, k)$-anonymity can solve the $d$-neighborhood attack of graph, or structure attack of graph easily, and make the node information loss is minimized according to $a$, $a$ can satisfy the user personalized needs simultaneously.

A $(d, k)$ Anonymity Algorithm

In Node Weighted algorithm the nodes with higher degree will be disposed first, and then the lower degree nodes will be calculated. The basic procedure of the algorithm is: (1) the initial graph G is converted to non-weight graph $G'$ based on the node weight $a (w_{ij})$; (2) the $d$-neighborhood lower triangular matrixes of all the nodes are computed based on $d$ neighborhood query encoding rules; (3) the nodes are sorted with the degree from large to small; (4) the nodes are anonymized and divided by the parameter $k$; (5) the anonymity graph including all the nodes are published.

The Step 1 is the initiation, and can be achieved in $O(1)$; the time complexity of Step 2 is $O(|V|^2)$, for all the nodes in the graph should be traversed; in Step 3, the neighborhood information of nodes are breadth first searched, so the time complexity is $O(|V|+|E|)$; the Step 4 with bubble sort is $O(|V|^2)$; from Step 5 to 22, the nodes are anonymized and divided, and the largest time complexity is $O(|V|^2)$. Above all, the time complexity of the whole algorithm is $O (1+|V|+|E|+ 2|V|^2) = O (|V|^2+|E|)$. In
undirected graphs, the number of edges is |E|, and the number of nodes is |V|. |E|≤|V|(|V|-1)/2, so the time complexity of a (d, k)-anonymity algorithm is O(|V|^2).

Experiments and Results

The FriendNet produced by Pajek is used for experiment data set. With Pajek, 5 undirected graphs with different order of magnitudes are created, and the numbers of nodes are respectively 5000, 10000, 15000, 20000, 25000.

Experiments and Results

The FriendNet produced by Pajek is used for experiment data set. With Pajek, 5 undirected graphs with different order of magnitudes are created, and the numbers of nodes are respectively 5000, 10000, 15000, 20000, 25000.

Figure 5. Nodes increasing percentage.

Figure 6. Edges increasing percentage.

The algorithms in this paper and paper [10] are compared in the aspect of information loss. Information loss can be measured amounts of adding new edges and adding new nodes. We use the method from paper [12]. The a (d, k)-anonymity algorithm in figure 5, figure 6, and figure 7 is named as NW algorithm, and the algorithm in paper [10] is named as HDVF algorithm. The experiments are accomplished in the PC with Intel Core(TM) 2 Duo CPUE4600 @ 2.40GHz, 2 GB memory, and the OS is Microsoft Windows XP. The programs are in VS.net 2010 IDE.

When the isomerized social network graphs changed to become isomorphic ones, a number of nodes and edges should be added in initial graphs. When the structure is more different, the number of adding is higher. Meanwhile, the information loss is larger.

Fig. (5) and Fig.(6) show these increasing conditions. Some nodes are added to construct the isomorphic graphs, the percentage of adding nodes in all the nodes of the graph is shown is Fig.(5), and the situation of edges is shown in Fig.(6).

Figure 8. Information loss degree when k=10.

Figure 9. Information loss degree when k=15.

In Fig.(7), Fig.(8) and Fig.(9), the information loss situation of HDVF and NW algorithm is shown with k=5, k=10 and k=15. Form these three figures, the difference of these two algorithms is small, and the information loss degrees are both decrease with the increasing of data scale. The reason is that the candidate set will be larger with the increasing of data scale, and finding similar neighborhood will be easier. Moreover, with the increase of d or k, the information loss will be larger. It is because...
with higher d or k value, it is difficult to get the neighborhood knowledge and the demand of privacy protection is higher. So the information loss degree is obviously larger.

Under normal conditions the information loss degree of NW algorithm is less than HDVF algorithm. In NW algorithm, social relationships are measured by node weights, and the social net are more complex and include more information, so the cost of anonymity is lower in larger data scale. For this reason, the information loss degree of NW algorithm is less.

Conclusions
In this paper, our main contributions are that we study the theory of k-anonymity of social network, try to use lower triangular matrix to represent the relationship among the nodes of social network, and propose a personalized a (d, k)-anonymity mode. The a (d, k)-anonymity algorithm experiments prove that it can make anonymous nodes effectively resist d-neighborhood attack of graph, and structure attack of graph, at the same time, make the node information availability according to the weighted parameter a given in advance, that can satisfy personalized needs.

Acknowledgement
This work was supported in part by “Application Technology Research and Development of Harbin” (Grant No. 2014RFQXJ062) and “Research Fund for the Innovative Scholars of Harbin” (Grant No.2016RAQXJ013).

Reference
[1] R. Kumar, J. Novak, A. Tomkins, “Structure and evolution of online social networks”, Proceedings of the 12th ACM SIGKDD international conference on Knowledge Discovery and Data Mining (KDD), pp.611-617, 2010.
[2] Y. J. Luo et al., “Overview of protecting user privacy in social networks”, Application Research of Computers, Vol.10, pp.3061-3064, 2010.
[3] L. Wu, M. Majedi, “Analysis of Social Networking Privacy Policies”, Proceeding of 2010 International Conference on Extending Database Technology/International Conference on Database Theory(EDBT/ICDT), pp.456-462, 2010.
[4] K. Liu, E. Terzi, “Towards identity anonymization on graphs”, Proceedings of the 2008 ACM SIGMOD international conference on Management of data, pp. 93-106, 2008.
[5] J. M. Han et al., “Research in Microaggregation Algorithms for k-Anonymization”, ACTA ELECTRONICA SINICA, vol. 36, no. 10, pp. 2021-2029, 2008.
[6] H. W. Luo, G. H. Liu, “(L, K)-anonymity based on clustering”, Journal of Yanshan University, vol. 31, no. 1, pp. 82-86, 2007.
[7] T. M. Truta, B. Vinay, “Privacy protection: p-Sensitive k-Anonymity Property”, In Proceedings of the 22nd International Conference on Data Engineering Workshops, pp. 94, 2006.
[8] B. Zhou, J. Pei, “The k-anonymity and l-diversity approaches for privacy preservation in social networks against neighborhood attacks”, Journal of Knowledge and information systems, Vol. 28, pp. 47-77, 2011.
[9] L. Zou, L. Chen, M. T. Oszu, “k-automorphism: General Framework for Privacy Preserving Network Publication”, Proceedings of the VLDB Endowment, vol.2, pp. 946-957, 2009.
[10] H. W. Wu, J. P. Zhang, B. Wang, et al., “(d, k)-anonymity for social networks publication against neighborhood attacks”, Journal of Convergence Information Technology, vol.8 (2): 59-67, 2013.
[11] H. W. Wu, J. P. Zhang, B. Wang, et al., “K+-isomorphism: privacy preserving publication against structural attacks in social networks”, International Journal of Advancements in Computing Technology, 4(22): 154-162, 2012.

[12] B. Zhou, J. Pei, “Preserving privacy in social networks against neighborhood attacks”, Proceedings of the 2008 IEEE 24th International Conference on Data Engineering, ICDE’08, pp 506-515, 2008.