

Privacy Preservation of Affinities in Social Networks

Lian Liu
HiPSCCS & CMIDA Lab,
Department of Computer
Science, University of
Kentucky, USA.
lliuc@csr.uky.edu

Jinze Liu
Department of Computer
Science, University of
Kentucky, USA.
liuj@cs.uky.edu

Jun Zhang
HiPSCCS & CMIDA Lab,
Department of Computer
Science, University of
Kentucky, USA.
jzhang@cs.uky.edu

ABSTRACT

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. So privacy preservation technologies should be exercised to protect social networks against various privacy leakages and attacks. Beyond the ongoing privacy preserving social network studies which mainly focus on node de-identification and link protection, this paper is written with the intention of preserving the privacy of link's affinities, or weights, in a finite and directed social network. To protect the weight privacy of edges, we define a privacy measurement, k -anonymous, over individual weighted edges. A k -anonymous weighted edge can make itself more indistinguishable from adjacent edges with respect to edge weights rather than node degrees. It is considered in this paper that modified weights of some edges should be released instead of the real ones to transform original weighted edges to k -anonymous edges, while preserving the shortest paths and the corresponding lengths between user-defined node pairs as much as possible. To achieve this goal, a probabilistic graph is used to model the weighted and directed social network. Based on this probabilistic graph, random walk, and matrix analysis, we present a modification algorithm on the weights of edges to accomplish a balance between the weight privacy preservation and the shortest path utilization. Finally, we give experimental results to support our theoretical analysis.

Categories and Subject Descriptors

H.2.7 [Database Management]: Database Administration—Security, Integrity, and Protection; H.2.8 [Database Management]: Database Applications—Data Mining

General Terms

Management, Security

Keywords

Social networks, Privacy, Probabilistic graphs

1. INTRODUCTION

The availability of digital technologies and internet development has promoted a proliferation of social networks. A social network consists of a set of entities with certain affinities (or weights) between them. Therefore, a social network can be simply represented by a graph, where each node corresponds to an entity and the weight of each edge between two nodes corresponds to an affinity. Both entities and edges might have attributes attached to them. For example, the attributes tied with nodes might include individual identification such as the Social Security Number (SSN) and personal features. The affinity is one of the major attributes of the edges in social networks which can deepen our understanding about the social network, such as the community evolution [14] and the modular structure [7]. Due to the public awareness of privacy protection, the sharing of certain social networks may be seriously hampered by the need for a balance between the protection of sensitive content and the public availability of data utility.

The research in privacy preserving social networks mainly focuses on the protection of node attributes, especially node's identification, via de-identification processes [2, 4, 5, 6, 11, 15, 17, 18, 19]. Recently, encouraged by the development of information sharing, there are an increasing number of applications where individual identification is not considered to be confidential. For instance, to facilitate research collaborations, some academic social networks, like ArnetMiner [14], allow the creation of the academic research network by mining bibliography databases and researchers' personal web sites through public web portals. In such academic social networks, nodes represent different researchers and edges typically represent the collaborative relationships between two researchers, such as the number of papers they collaborate on. Such networks typically are useful in querying related research conducted by different researchers. In this case, privacy of individual researchers is not a major concern given these networks are constructed from the public data. On the other hand, it is worthwhile noting that though legal, such networks derived from public databases make implicit affinities between researchers much more explicit and specific, and consequently, invading the privacy of

individual researchers.

Therefore, beyond the ongoing privacy preserving social networks which mainly focus on node de-identification and link protection, the significance about the privacy of edge weights also deserves to be seriously studied.

Although various de-identification techniques can always be applied to hide the affinity information from third parties, we argue in this paper that de-identification not only violates the increasing need for information sharing but also seriously limits the utility of the weighted social network, such as modular structure [7] which is stored in the weights rather than the graph node topology.

Our goal is to maximize both information sharing and data utility while at the same time preserving privacy. In this paper, the *data utility* of a weighted social network is concerned with the disclosure of the shortest paths and their lengths in this network. The *data privacy* about weights is related to the discrepancy between the weights of adjacent edges. The discrepancy related to weighted edges should make themselves more indistinguishable from their adjacent edges with respect to edge weights.

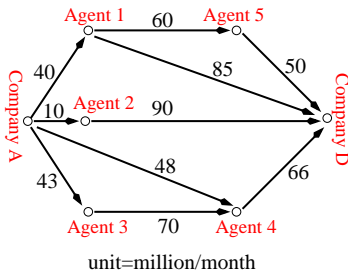


Figure 1: A Weighted and Directed Social Network.

The following example illustrates the importance of the balance between data utility and data privacy. Assume there is a conceptual business social network, derived from an automaker business network between Japanese automotive corporations and American suppliers in North America [8]. The network is a weighted and directed social network as shown in Figure 1, where each node represents a business agent, and each weighted edge represents the cost amount from one business unit to another. The affinities (or the costs) in the business network are important in providing useful global information such as the optimal supply chain (a path with the lowest cost between companies).

In the example, the optimal supply chain typically exists between Japanese automotive corporations and their first-tier Japanese suppliers such as Agent 1, Agent 5, Agent 2, and Agent 4. However, due to the sensitivity of importing cost and potential political pressure against mass outsourcing, Japanese suppliers are encouraged to collaborate with qualified local North American suppliers. Many North American suppliers, therefore, compete to become a subcontractor of Japanese corporations. In this case, if exact transaction information is released along the optimal supply chain, other companies might offer a new quotation below the existing costs in order to win the contract.

Therefore, in Figure 1, the data utility is the shortest path of transaction expenses between Company A and Company D and its length. But the weight privacy in the example is not good since the discrepancy between some edges, like the edge (Company A→Agent 2) is much different from the

three others. It is easy for business adversaries to breach the privacy of the edge (Company A→Agent 2) if some background information, such as one edge weight is much far away from the other weights, is publicly available.

To accomplish it, we reduce a weighted graph into a probabilistic graph. A probabilistic graph is a general graph model where the transition probability from one node to the others defines the affinities between two identities. Although replacing the original weights by probabilities is a privacy preserving approach to some extent, the privacy concerned in this paper goes well beyond this process.

Following the work on the weight privacy preserving social network [12] that gives a weight preservation scheme based on the network topology and a greedy algorithm, this paper is written with the intention of publishing a privacy preserving social network in which we only modify edge weights to minimize the weight discrepancy without adding or deleting any node and edge, while the shortest paths and the corresponding lengths between user-defined node pairs in the modified social network are maintained to be as close to the original ones as possible.

Our contributions in this paper are summarized as follows:

1. We construct a probabilistic graph which can inexpensively perform a quantitative analysis on data utility and data privacy.
2. A k -anonymous privacy is defined to measure the privacy level of individual edge weights. We propose a k -anonymous privacy over the continuous weights since the standard k -anonymity, defined on discrete and categorical values [13], is not applicable to continuous values.
3. Based on our proposed single edge weight modification algorithm and the quantitative analysis of the modification, we construct an edge frequency order to achieve the balance between data utility and data privacy.

The remaining parts of this paper are organized as follows. A brief survey of related work regarding privacy preserving social networks is given in Section 2. The definition and discussion about data utility and data privacy are presented in Section 3. The detailed procedure of privacy preserving weight modification is described in Section 4. Experimental results and a conclusion are discussed in Section 5 and Section 6, respectively.

2. RELATED WORK

In addition to a large amount of privacy preserving data mining literature, more and more researchers have paid their attention to preserving privacy of social networks. This section provides a brief survey on privacy preserving social networks. The literature on general privacy preserving data mining without emphasis on social networks, such as variants of tabular k -anonymity and distributed privacy preserving data mining, is excluded from this survey.

Backstrom et al. [2] described a framework to distinguish the possibility of a certain edge existed in a social network. It shows that the identification of almost any node is easy to be leaked based on the implantation. Korolova et al. [9] developed a breach analysis on the node's identification just based on a part of background information regarding the neighborhood.

Based on these theoretical analysis, researchers developed various algorithms to add/delete some edges to break the chances of differentiating the given nodes and/or edges from de-identification social networks. They placed emphases on the protection of social entity's identification via de-identification k -anonymity and variants. For example, Wang et al. proposed a logic function to quantify the node anonymity in [15]. Hay et al. [5, 6], Zhou et al. [19], and Liu et al. [11] presented an essentially similar scheme to add and/or delete some unweighted edges in social networks to keep malicious users from accurately re-identifying target nodes based on auxiliary information about the number of neighbors. Cormode et al. [4] gave a bipartite anonymity method to group sensitive nodes into an aggregate class via a safe-group technique. Ying et al. [17] discussed the relationship between the ability to breach the edge identification and the degree of edge randomization from the viewpoint of eigenspace. Acquisti et al. [1] claimed a different case in which they incorporated publicly available information into the privacy preserving social network to breach personal information. Zheleva et al. [18] hid and removed some edges based on edge clustering methods in an edge-labeled model in which unweighted edges are considered to be confidential. Interested readers can refer to [10] for a comprehensive discussion about privacy preserving social networks against the disclosure of confidential nodes and links.

These methods all focus on preserving either node or edge privacy. Although Liu et al. [12] proposed weight privacy preservation in social networks, their algorithm depends on the topology information and only gives a range for the weight modification without an explicit privacy guarantee. In this paper, we emphasize on modification of edge weights in accordance with the requirement of privacy protection and data utilization. In other words, data owners are not willing to disclose real weights of edges, but would like to keep some shortest paths and the corresponding lengths for the data analysis purpose.

3. DATA UTILITY AND PRIVACY

In this section, a detailed weight modification in accordance with the data utilization and the privacy preservation will be given. We first give some preliminaries and notations that will be used later.

We define a social network in this paper as a weighted and directed graph $G=\{V, E, W\}$. The nodes of the graph, V , are abstract representation of any meaningful entities. Here, we do not de-identify nodes of social networks, especially in identification-public social networks such as academic collaboration networks. E is the set of all directed and weighted edges. One positive quantitative weight, $w_{i,j}$, between node i and node j , is tied to the directed edge which reflects the affinity between the two entities. And we assume that all weights in this paper are positive. If there is no edge between two nodes, the corresponding weight is denoted as a large enough number. The adjacency matrix, W , of the social network is composed of all edge weights $w_{i,j}$. The shortest path between two different nodes is a path whose total sum of the weights of the passing edges is the smallest one among all possible paths. We assume the cardinalities of V and E , $n=||V||$ and $m=||E||$, are the numbers of nodes and edges in this social network, respectively. Although our algorithm is based on directed graphs, it can be easily extended to undirected graphs. Each undirected and weighted

edge can be transformed into two directed edges between the same node pair with opposite directions and the same weights while the graph topology is unchanged. We assume that $R_{i,j}$ is the set of all possible paths connecting node i and node j , and $r_{i,j}$ is a particular path from node i to node j . R and r are short for $R_{i,j}$ and $r_{i,j}$ without otherwise explicitly stated.

We assume in the publication model of privacy preserving social networks that each party in this social network owns a local private network, and submits its local network to a trusted third-party central processor which will not collude with anyone. After collection, the central processor modifies weights of all edges according to privacy and utility requirements. Then it publishes the modified global social network to the public for future privacy preserving social network data mining. The published modified global social network is denoted as $G^*=\{V^*, E^*, W^*\}$. The second publication model is that one data owner has the entire social network and submits it to the trusted third-party processor for the same privacy preserving processing.

3.1 Data Utility

Before our formal definition about the probabilistic graph, the adjacent edge set $\Phi(i)$ of a given edge $(i \rightarrow j)$ is defined as $\Phi(i)=\{the\ edge\ (b \rightarrow c) \mid b = i \ \& \ w_{b,c} \neq 0\}$. Intuitively, the adjacent edge set $\Phi(i)$ is an edge set in which all edges come from the same source node i in the graph. Let γ_i be the cardinality of $\Phi(i)$.

The weight in weighted graphs is transformed into a transition probability as in Definition 1, based on the adjacency matrix W of a social network.

Definition 1. The transition probability, $p_{i,j}$, of a given directed and weighted edge $(v_i \rightarrow v_j)$ is defined as

$$p_{i,j} = \frac{1}{\sum_{t=1}^{\gamma_i} \frac{w_{i,j}}{w_{i,t}}}. \quad (1)$$

Intuitively, an edge with a small weight is more likely to be chosen as a part of the shortest path correspondingly. Since $p_{i,j}$ is inversely related to the weight, based on Definition 1, an edge with a large $p_{i,j}$ is more likely to be chosen as an edge in the shortest path than the one with a small $p_{i,j}$.

The shortest path from node i to node j in weighted graphs is equivalent to a path r whose probability $P(r)$, defined in Formula (2), is highest among all possible paths between the two nodes.

$$P(r) = \frac{\exp[-\theta E(r) + \ln \bar{P}(r)]}{Z_{i,j}}, \quad (2)$$

where, $\bar{P}(r) = \prod_{t=1}^{\tau(r)} p_{v_t, v_{t+1}}$, $E(r) = \sum_{t=1}^{\tau(r)} w_{v_t, v_{t+1}}$, $Z_{i,j} = \sum_{r \in R} \exp[-\theta E(r) + \ln \bar{P}(r)]$, $\tau(r)$ is the number of edges in the path r , and θ is a parameter, say 20 [16]. The $E(r)$ is the sum of edge weights in the path of weighted graphs, and $\bar{P}(r)$ is the product of edge transition probabilities in a path. Formula (2) implies that the shortest path has the highest probability $P(r)$ among all possible paths. Moreover, the smaller $E(r)$ a path has, the higher $P(r)$ it is.

The length of the shortest path between node i and node j in weighted graphs is translated into the expected energy, \bar{E} , defined as follows:

$$\bar{E} = \sum_{r \in R_{i,j}} \frac{\exp[-\theta E(r) + \ln \bar{P}(r)] E(r)}{Z_{i,j}}. \quad (3)$$

Here, $E(r)$ is the sum of edge weights for any path r (r is not required to be the shortest path), and \bar{E} is the length of the shortest path.

Due to the page limitation, we omit the detailed introduction for the derivation procedures of Formulas (2) and (3). Interested readers can consult paper [16] for a comprehensive insight.

From the viewpoint of probabilistic graphs, the shortest path is a path with the highest probability $P(r)$ and the corresponding length being \bar{E} . To calculate the possibility of a given path, the numerator of Formula (2), $\exp[-\theta E(r) + \ln \bar{P}(r)]$, is easy to calculate given the path is known. But the computation of the denominator $Z_{i,j} = \sum_{r \in R} \exp[-\theta E(r) + \ln \bar{P}(r)]$ is hard since it requires to enumerate all possible paths. The difficulty in computing \bar{E} is similar. Alternatively, the computation of $Z_{i,j}$ can be transformed into the computation of matrix power series. Before discussing the computation of $Z_{i,j}$, a matrix Q is defined as:

$$Q = \exp[-\theta \tilde{W} + \ln \tilde{P}], \quad (4)$$

where, \tilde{W} is the same as W with the exception of the j -th row of \tilde{W} being positive infinite (in practice we use a very large positive number), and \tilde{P} is a matrix composed of $p_{i,j}$. But the j -th row of \tilde{P} is 0. For example, the (i, j) -th entry of Q^3 (or $Q * Q * Q$) equals to $Z_{i,j} = \sum_{r \in R(3)} \exp[-\theta E(r) + \ln \bar{P}(r)]$, where $R(t)$ denotes a set of paths connecting node i and node j by t edges.

Based on the matrix Q , $Z_{i,j}$ can be computed as:

$$\begin{aligned} Z_{i,j} &= \sum_{r \in R_{i,j}} \exp[-\theta E(r) + \ln \bar{P}(r)] \\ &= \sum_{t=1}^{\infty} \sum_{r \in R_{i,j}(t)} \exp[-\theta E(r) + \ln \bar{P}(r)] \\ &= \sum_{t=1}^{\infty} Q^t. \end{aligned} \quad (5)$$

Under the condition $i \neq j$ and the absolute value of maximal eigenvalue of Q is smaller than 1, $Z_{i,j} = \sum_{t=1}^{\infty} [Q^t]_{i,j} = [(I - Q)^{-1} - I]_{i,j} = e_i^T (I - Q)^{-1} e_j$, where $[\cdot]_{i,j}$ denotes the (i, j) -th entry of the matrix and e_i is the i -th column of an identity matrix with proper dimension. Here, the computation for the sum of possibilities for all paths can be transformed into computing the inverse of a matrix.

Similarly, the length of the shortest path, \bar{E} , is calculated as $\bar{E} = -\frac{z_i^T * S * z_j}{Z_{i,j}}$ (please refer to [16] for detail), where, z_i , z_j and $Z_{i,j}$ are the i -th, j -th columns, and the (i, j) -th entry of the matrix Z , and $S = \exp[-\theta \tilde{W} * \ln \tilde{W} + \ln \tilde{P} * \ln \tilde{W}]$.

Until now, from the viewpoint of probabilistic graphs, the shortest paths and the corresponding lengths between node pairs are introduced. We will propose the modification scheme to approach a balance between privacy preservation of edge weights and utilization of the shortest path in the next subsection.

3.2 Data Privacy

One of the two purposes in this paper is to protect the weight privacy of the edges. Intuitively, an edge with an indistinguishable weight is relatively difficult to breach based on the background information about adjacent edges in this social network.

For example, Company A has four directed edges to Agent 1, Agent 2, Agent 3, and Agent 4 with corresponding weights 40, 10, 48, and 43 as in Figure 2(a). Intuitively, it is possible for us to guess that the edge (Company A \rightarrow Agent 2) is the one if we have background information such that one weight is far more different than the others. But if the weights of the four edges are very close to each other, like 35, 32, 36, and 33 as in Figure 2(b), it is not easy to know which one is the distinguishable edge. Also note that in the modified network as in Figure 2(b), the shortest path between Company A and Company D (Company A \rightarrow Agent 2 \rightarrow Company D) is the same as the original one in Figure 2(a), and the corresponding length (99) in the modified network is very close to the original one (100) in Figure 2(a).

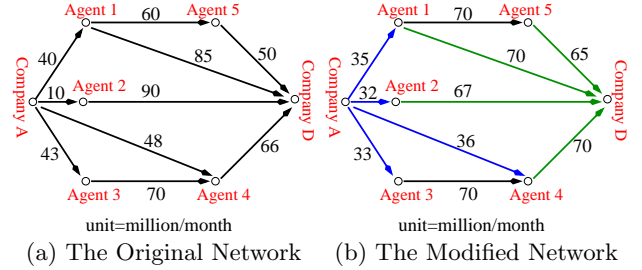


Figure 2: The original business social network and the modified one. In the modified network, the blue edge group and the green edge group satisfy the 4-anonymous privacy where $\mu=10$.

To eliminate the distinguishability between edge weights, we propose the definition of a k -anonymous weight privacy as follows:

Definition 2. The edge ($i \rightarrow j$) is k -anonymous if and only if there exist at least k edges in $\Phi(i)$ whose weights w_{i,t_l} , $l=1, \dots, c$, and $c \geq k$, satisfy $\|w_{i,j} - w_{i,t_l}\| \leq \mu$, $l=1, \dots, c$.

Here, μ is a predefined positive parameter to control the degree of privacy and $\Phi(i)$ is the adjacent edge set in which all edges come from the i -th node. Please note that in the case of the total number of edges in $\Phi(i)$ being smaller than k , we still think the edge is k -anonymous if all edges in $\Phi(i)$ are not far away more than μ .

In Figure 2(a), the shortest path between Company A and Company D is the path (Company A \rightarrow Agent 2 \rightarrow Company D), and the corresponding length is 100. But the privacy of the edges in the original network is not good since many adjacent edges are not indistinguishable, i.e., the edge (Company A \rightarrow Agent 2) is much different from the three others, the edge (Agent 1 \rightarrow Company D) has a big difference from the edge (Agent 1 \rightarrow Agent 5), and do so the four incoming edges to Company D. After modification, as in Figure 2(b), most edges are indistinguishable from their adjacent edges as both the blue edge group and the green edge group satisfy a 4-anonymous privacy, where $\mu=10$. At the same time, the shortest path in the modified network is the same as the one in the original network, while the corresponding modified length is 99 and the original one is 100.

From the perspective of a probabilistic graph, Definition 2 is equivalent to the following definition.

Definition 3. The edge $(i \rightarrow j)$ is k -anonymous if and only if there exist at least k edges in $\Phi(i)$ whose transition probability p_{i,t_l} , $l=1, \dots, c$, and $c \geq k$, satisfy $\|1/p_{i,j} - 1/p_{i,t_l}\| \leq \mu\Delta$, $l=1, \dots, c$ and $\Delta = \sum_{l=1}^c 1/w_{i,t_l}$.

Formally, we use the following theorem to decide whether an edge is k -anonymous or not.

THEOREM 1. An edge $(i \rightarrow j)$ is k -anonymous if

$$\sum_{l=1}^{\gamma_i} \text{sgn}(\| \frac{1}{p_{i,j}} - \frac{1}{p_{i,t_l}} \| - \mu\Delta) \leq \gamma_i - k.$$

Here, $\text{sgn}(\cdot)$ is a modified sign function such that

$$\text{sgn}(x) = \begin{cases} 1 & x \geq 0, \\ 0 & \text{otherwise.} \end{cases}$$

Theorem 1 can be proved by straightforwardly using Definition 3. The meaning of Theorem 1 is that an edge $(i \rightarrow j)$ is not k -anonymous if the number of edges whose weights are more than μ far away from that of this given edge is larger than $\gamma_i - k$.

Measured over the weighted edges, our k -anonymous privacy definition has a substantial difference from the anonymity privacy definition on nodes [5, 6, 9, 10, 11, 19]. As mentioned in the introduction, the privacy against the disclosure of node information is unnecessary in some cases. Furthermore, previous node privacy definitions such as k -anonymous nodes [5, 6] and neighborhood attacks [2, 19] are not easy to be extended to the confidential weights of the edges since edge weights are continuous values and the node privacy definitions are almost essentially based on discrete node degrees.

Although we hope to construct a social network with any amount of edges to achieve the same data utility and a k -anonymous privacy, it is not always possible to do so due to the shortest-path data utilization being a strong constraint when the number of the shortest paths to be maintained is large. So, our privacy preserving purpose is to make as many edges k -anonymous as possible.

4. MODIFICATION ALGORITHM

Although there is at least one shortest path between any pairs of nodes in a connected graph, it is reasonable to assume that not all shortest paths are equally important. In addition, it has been proved that it is impossible to modify each weight and preserve all the shortest paths and the corresponding lengths [12]. We assume that the data owners decide about the subset of all shortest paths to preserve. Our task is, given a set of targeted shortest paths H , to maximize both the weight privacy preservation and the shortest path utilization.

We will present the algorithm for the single edge modification in Section 4.1, and will detail the method to choose an optimal order to modify multiple edges in Section 4.2.

4.1 Single Edge Weight Modification

The change of a single edge weight can affect both the shortest paths passing through it and not passing through it. To modify the weight $w_{i,j}$ of a given directed and weighted edge $(i \rightarrow j)$ without changing the set of the shortest paths in H , several conditions needed to be satisfied and they are listed in Figure 3.

1. $V^* = V$ and $E^* = E$,
2. $P(r^*) > P(r)$, for each shortest path r in H where edge $(i \rightarrow j)$ is in r ,
3. $P(r^*) < P(r)$, for each shortest path r in H where edge $(i \rightarrow j)$ is not in r ,
4. $E(r^*) \approx E(r)$, for each shortest path r where edge $(i \rightarrow j)$ is in r ,
5. $\sum_{l=1}^{\gamma_i} \text{sgn}(\| \frac{1}{p_{i,j}} - \frac{1}{p_{i,t_l}} \| - \mu\Delta) \leq \gamma_i - k$.

Figure 3: The conditions of weight modification of a single edge.

Condition 1 implies that the topology of the social network (node structure $V=V^*$ and edge structure $E=E^*$) will not be changed. Conditions 2 and 3 make sure that after the modification, the shortest paths in the target set H are still the shortest paths and a non-shortest path is not likely to become a shortest path. Condition 4 states that we need to maintain not only the shortest paths, but also their lengths. Condition 5 says that the weight $w_{i,j}$ of the edge should be modified so that it becomes a k -anonymous edge as much as possible.

In Algorithm 1, we summarize the steps to determine a modification value e with respect to the weight $w_{i,j}$ for the satisfaction of conditions in Figure 3. Several inequalities need to be solved in order to find a potential new weight that satisfies the above constraints. These inequalities include Formulas (6) and (7). Solving these inequalities together will possibly result in a feasible range where the modification value e can be selected from. Then the best e within the range which minimizes the k -anonymous privacy will be selected.

Algorithm 1 Single Edge Weight Modification Algorithm.

Input: The weight $w_{i,j}$ of the edge $(i \rightarrow j)$ and the social network $G=(V, E, W)$, and the set H of the selected shortest paths to be maintained.

Output: The modification value e with respect to $w_{i,j}$.

- 1: Initialize $U_1=(-\infty, +\infty)$
- 2: **for** each path r in H **do**
- 3: **if** edge $(i \rightarrow j)$ is in r **then**
- 4: solve the inequality, and let its answer be U'

$$P(r^*) = \frac{\exp[-\theta E^{new}(r^*) + \ln \bar{P}^{new}(r^*)]}{Z_{i,j}^{new}} > P(r) \quad (6)$$

- 5: **else**
- 6: solve the inequality, and let its answer be U'

$$P(r^*) = \frac{\exp[-\theta E^{new}(r^*) + \ln \bar{P}^{new}(r^*)]}{Z_{i,j}^{new}} < P(r) \quad (7)$$

- 7: **end if**
- 8: $U_1=U_1 \cap U'$.

9: **end for**

- 10: **if** $U_1 \neq \emptyset$ **then**

11: $U_2=\{e \mid e \in U_1 \ \& \ \bar{E}^{new} \approx \bar{E}\}$

12: **else**

13: EXIT

14: **end if**

15: $e=\operatorname{argmin}_{e \in U_2} \sum_{l=1}^{\gamma_i} \text{sgn}(\| \frac{1}{p_{i,j}} - \frac{1}{p_{i,t_l}} \| - \mu\Delta)$

In the following, we show how to calculate the Formulas (6) and (7) when the weight is modified from $w_{i,j}$ to $w_{i,j}^* = w_{i,j} + e$.

When a weight is updated, its corresponding new transition probability $p_{i,j}$ can be computed based on Definition 1 as follows:

$$p_{i,j}^{new} = \frac{\frac{1}{w_{i,j}+e}}{\sum_{t=1 \& t \neq j}^{\gamma_i} \frac{1}{w_{i,t}} + \frac{1}{w_{i,j}+e}} \quad (8)$$

Formula (8) implies the following information. Firstly, since all weights are positive, the value of e should be larger than $-w_{i,j}$. Secondly, the change of one edge weight is only effective in $p_{i,j}$ and has nothing to do with other probabilities. Thirdly, Formula (8) is a monotonically decreasing function with respect to e in the range $(-w_{i,j}, +\infty)$ by rewriting it into the form of $p_{i,j}^{new} = \frac{1}{\delta*(w_{i,j}+e)+1}$ (here, $\delta = \sum_{t=1 \& t \neq j}^{\gamma_i} \frac{1}{w_{i,t}}$ is a constant in the case of only $w_{i,j}$ being changed). This monotonically decreasing property means that the probability $p_{i,j}^{new}$ will be increasing as long as $w_{i,j}$ is decreasing and the vice versa.

The new probability of the shortest path as referred to in Formulas (6) and (7) concerns both $E^{new}(r^*)$, $P^{new}(r^*)$ and $Z_{i,j}^{new}$. Here we focus on the computation of $Z_{i,j}^{new}$ since the rest are straightforward to compute ($E^{new}(r^*) = E(r) + e$, $P^{new}(r^*) = P(r) * \frac{p_{i,j}^{new}}{p_{i,j}}$).

We will discuss how to calculate the numerator and denominator of $P(r^*)$ as in Formulas (6) and (7) in the following paragraphs from the viewpoint of matrix perturbation.

Regarding the numerator, $\exp[-\theta E^{new}(r^*) + \ln \bar{P}^{new}(r^*)] = \exp[-\theta(E(r) + e) + \ln(\bar{P} * \frac{p_{i,j}^{new}}{p_{i,j}})]$.

With respect to the denominator, the updating algorithm for Q is proposed first since $Z_{i,j}$ is related to Q according to Formula (5). The new Q , denoted as Q^{new} , is reconstructed as:

$$\begin{aligned} Q^{new} &= \exp[-\theta \widetilde{W} + \ln \widetilde{P}^{new}] \\ &= \exp[-\theta(\widetilde{W} + e) + \ln(\widetilde{P} * \frac{p_{i,j}^{new}}{p_{i,j}})] \\ &= \exp(-\theta e) \frac{p_{i,j}^{new}}{p_{i,j}} Q, \end{aligned} \quad (9)$$

Here, θ is a user-defined parameter and e is the value for the weight modification, $p_{i,j}^{new}$ is the updating transition probability of the modified weight $w_{i,j}^*$ as in Formula (8), $p_{i,j}$ is the transition probability of the weight $w_{i,j}$ in the original social network as in Formula (1), and Q is the original value as in Formula (4). Because $Q/p_{i,j}$ in Formula (9) is only related to the original social network, we can consider it as a constant.

$Z_{i,j}^{new}$ is the (i,j) -th entry of the matrix $(I - Q^{new})^{-1}$. So, based on Formula (9),

$$Z_{i,j}^{new} = (I - Q^{new})^{-1} = (I - Q + [\beta])_{i,j}^{-1} \quad (10)$$

$$= (I - Q)^{-1} - \frac{\beta}{1 + \beta Z_{j,i}} z_i z_j^T \quad (11)$$

$$= Z_{i,j} - \frac{\beta}{1 + \beta Z_{j,i}} Z_{i,j}. \quad (12)$$

Here, $[\beta]$ is a one-entry matrix with the (i,j) -th entry being $[1 - \exp(-\theta e) \frac{p_{i,j}^{new}}{p_{i,j}}]Q$. $Z_{i,j}$, $Z_{j,i}$, z_i and z_j are the (i,j) -th entry, the (j,i) -th entry, the i -th column and the j -th row

of the matrix Z . Because the four items have nothing to do with the modification and are known to the data owner, they can be computed in the preprocessing step to reduce the computational cost. The derivation from Formula (10) to Formula (11) is based *Sherman-Morrison-Woodbury formula*, $(A + UCV)^{-1} = A^{-1} - A^{-1}U(C^{-1} + VA^{-1}U)^{-1}VA^{-1}$. Particularly, based on results in [3], when the perturbation matrix UCV is a one-entry matrix D , the inversion of the perturbed matrix is as $(A + D)^{-1} = A^{-1} - \bar{A}^{-1}(1 + \bar{D}\bar{B})^{-1}\bar{D}\bar{B}$, where A is an $n*n$ positive matrix, D is a one-entry matrix with the (i,j) -th entry being a non-zero number e , \bar{A}^{-1} is the i -th column of A^{-1} , \bar{D} is e , \bar{B} is the (j,i) -th entry of A^{-1} , and \hat{B} is the j -th row of A^{-1} .

A new length of the corresponding shortest path is decided by W^* and Z^{new} 's components (the i -th, the j -th columns and the (i,j) -th entry of Z^{new}). So, the new length of this shortest path is computed as:

$$\bar{E}^{new} = \frac{z_i^T(\widetilde{W} + [e]) \circ (\widetilde{P} \circ [\frac{p_{i,j}^{new}}{p_{i,j}}]) \circ \exp[-\theta(\widetilde{W} + [e])]z_j^T}{z_{i,j}^{new}}. \quad (13)$$

Here, the operator \circ is the elementwise matrix multiplication, and z_i , z_j and $z_{i,j}^{new}$ are the i -th, the j -th columns and the (i,j) -th entry of Z^{new} . $[\frac{p_{i,j}^{new}}{p_{i,j}}]$ and $[e]$ are one-entry matrices with the corresponding values, respectively. The computation of z_i , z_j and $z_{i,j}^{new}$ is identical to Formula (12).

Therefore, to satisfy Condition 4 in Figure 3, \bar{E}^{new} should be close to \bar{E} which is a fixed value and known to the data owner. Note that if e throughout the computation of Formula (13) is not in the range of Formulas (6) and (7), we should discard this value and choose a bound value in the range. If an optimal weight modification is impossible to choose due to the boundary limitation in Formulas (6) and (7) at one step, we still can obtain a close length of the shortest path in the modification process of other weights.

4.2 Multi-Edge Modification Order

Although all edges can be randomly selected for modification, different orders of modification do not give the same level of privacy. We discuss a special order to modify the set of edges in order to achieve a high k -anonymous privacy while maintaining the same data utilities. Note that each edge weight is modified only once.

Decreasing an edge weight will increase the probabilities of the paths containing this edge while increasing an edge weight will decrease their probabilities. These were shown in Lemma 1.

Lemma 1.

$$P(r^*) = \frac{\exp[-\theta E^{new}(r^*) + \ln \bar{P}^{new}(r^*)]}{Z_{i,j}^{new}}$$

increases for a negative e and decreases for a positive e . Here $E^{new}(r^*)$, $P^{new}(r^*)$ and $Z_{i,j}^{new}$ are the functions of e .

Decreasing the weight of a given edge has two consequences: (1) The probability of the shortest paths going through this edge will increase, i.e., they are still the shortest paths; (2) The probability of the non-shortest paths going through this edge will also increase. It is possible that they become the shortest paths since their new probabilities are increasing. Therefore, there exists a range of the modification value e such that, after modification, the shortest

paths will stay the same, and the non-shortest paths will not become the shortest paths. From the perspective of the probability of the shortest paths, based on Lemma 1, we can make the range of the weight modification value e explicit in the following.

THEOREM 2. *For a given edge, the quantitative range U' of e in the Formulas (6) and (7) in Algorithm 1 is determined by the following range*

$$P(r_1)(Z_{i,j}Q - \frac{Z_{i,j}}{1 + QZ_{j,i}}) \leq \exp(-\theta e) \frac{p_{i,j}^{new}}{p_{i,j}} \leq P(r_2)Z_{i,j}Q,$$

where $P(r_1) = \max\{P(r) \mid \text{the edge is not in the shortest path } r \text{ of } H\}$, $P(r_2) = \min\{P(r) \mid \text{the edge is in the shortest path } r \text{ of } H\}$, the definitions of Q , Z , $p_{i,j}^{new}$ and $p_{i,j}$ are in Formulas (4), (5), (8), and (1).

Due to the page limitation, we omit the proof details of Lemma 1 and Theorem 2.

From Theorem 2, we can conclude that the range of the weight modification value e for a given edge is affected by the number of the shortest paths going through it, i.e., the more shortest paths go through this edge, the tighter the range of the weight modification is.

Although the modification range for a high frequency edge is tight, the k -anonymous privacy can probably be achieved by the weight modification of low frequency edges which have a bigger range and are modified later. Based on the observation, all edges are sorted in terms of their presence frequencies in the shortest paths. We first modify the weight of one edge whose presence frequency in the shortest paths is highest. Such sorting can achieve a high k -anonymous privacy.

5. EXPERIMENTAL RESULTS

One real database, EIES (Electronic Information Exchange System) Acquaintanceship at time 2, and two synthetic databases will be used for experiments.

The social network in the EIES dataset is a directed and weighted graph in which the data were collected to measure the acquaintanceship between 48 researchers to show their cooperation in research activities. In addition to the EIES database, to test the efficiency and scalability of our algorithm, we created two synthetic databases, SYN1 and SYN2. SYN1 is a social network with 100 objects in which every node is connected to each other and the weight is randomly selected from 10 to 100. SYN2 consists of 200 objects and 70% objects are connected with each other, and the weights of the edges range randomly from 10 to 100. Its corresponding weight matrix W is a 200*200 nonsymmetric matrix.

Comparison about the modification orders. We first show that our order of weight modification is better than other orders including the random one.

In Figures 4(a), 4(b) and 4(c), the y -axis is the percentage of k -anonymous edges, and x -axis is the different values of k . These figures show that our frequency sorting can achieve a higher level of k -anonymous privacy compared to other two sortings in all three social networks with different values of k .

Comparison about different sizes of H . The efficiency of the edge weight modification is really dependent on the ratio of the size of H to all node pairs in a social network. The more shortest paths and the corresponding

lengths we need to preserve, the less room of improvement we can achieve. So we choose several different sizes of H such as 5%, 10%, 15%, 20%, and 25% of all nodes pairs in order to test our algorithm. All the node pairs in H are randomly selected. The parameter θ is chosen as 20.

The purpose of our experiments is to show three things. 1). The ratio of k -anonymous edges to all edges. 2). The percentage of the shortest paths with respect to the node pairs of H in the modified social network is the same as the real one in the original social network. 3). The ratio of the length difference between the modified shortest path and the original one to the length of the original shortest path. The first criterion denotes the degree of weight privacy preservation, and the second and third criteria stand for the shortest path utilization.

In Figure 5(a), at x -axis 0.15, the circle line point is 0.94 (94%) and the star line point is 0.9, and the marked line point is 0.21. It means that, after our modification scheme, 90% edges are k -anonymous, 94% of the shortest paths of the node pairs in H is the same as the real ones in the original social network, and the relative difference between the lengths of the original shortest paths and that of the modified ones is 0.21, i.e., $\sum_{i \neq j \& (i,j) \in H} \frac{\|\bar{E}_{i,j}^{new} - \bar{E}_{i,j}\|}{\bar{E}_{i,j}} = 0.21$.

From Figures 5(a), 5(b) and 5(c), the circle line is high and smooth in all three figures. It means that most modified shortest paths are able to be kept the same as the real ones even if a large amount (40%) of node pairs in H need keep exactly the same shortest paths and close shortest path lengths. The more information we need to maintain (the size of H is increasing), the less privacy we can improve (the ratio of weight modification is decreasing). But the ratio is still large (they are all around 80% at x -axis 0.25). We point out that in the three original social networks, the percentages of k -anonymous edges are 62%, 42% and 49%. After modification, however, the percentages of k -anonymous edges increase to an average of 80% which means that our scheme still brings about a big hindrance for the weight privacy breach compared to the original level of privacy.

Comparison about different k . Figures 6(a), 6(b), 6(c) show the weight privacy in terms of the percentages of k -anonymous edges with different values of k .

In Figures 6(a), 6(b), 6(c), the green circle line denotes the ratio of the number of k -anonymous edges to that of all edges in the original social networks, and the blue star line means the k -anonymous edge ratio in the modified one. We can see that both the privacy level (circle line) in the original network and the privacy (star line) in the modified social network are decreasing as the value of k is increasing. But there are remarkable privacy differences between all original social networks and the corresponding modified networks. It demonstrates that our scheme can definitely increase the privacy preservation of original social networks to a noticeably higher level in different privacy protection requirements such as various k . More importantly, the preservation probability of the shortest paths (square line) are still maintained at a smooth level since we first keep the shortest-path utility before the data privacy. Although the lengths of the shortest paths (cyan diamond line) in modified social networks increase as k increases, the slope is not so sharp as the corresponding lines (red mark line) in Figures 5(a), 5(b), and 5(c). It implies that the relative difference between the lengths of the original shortest paths and that of modified

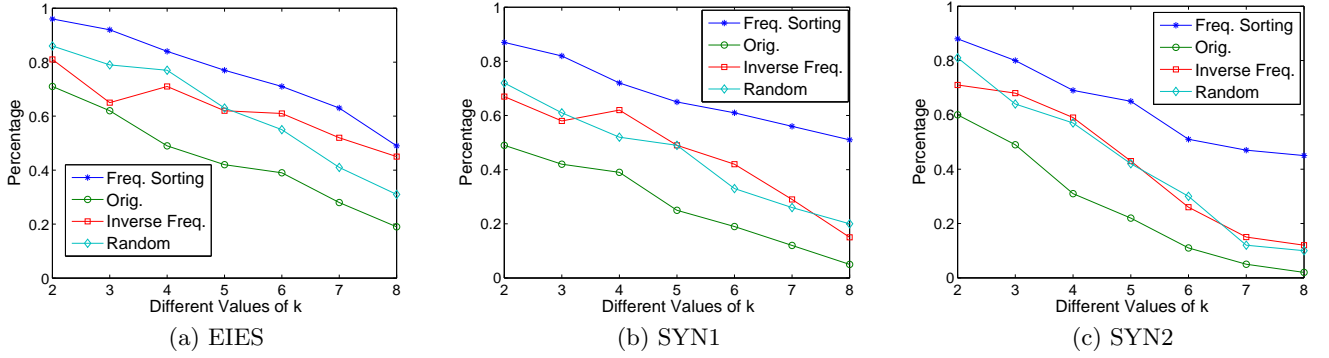


Figure 4: The comparison about privacy levels in three sortings and the original case in the condition of $H=10\%$ and $\mu=10$. The Freq. Sorting is the descending frequency sorting, the Orig. is the privacy level of the original social network, the Inverse Freq. is the ascendent frequency sorting, the Random is one random sorting.

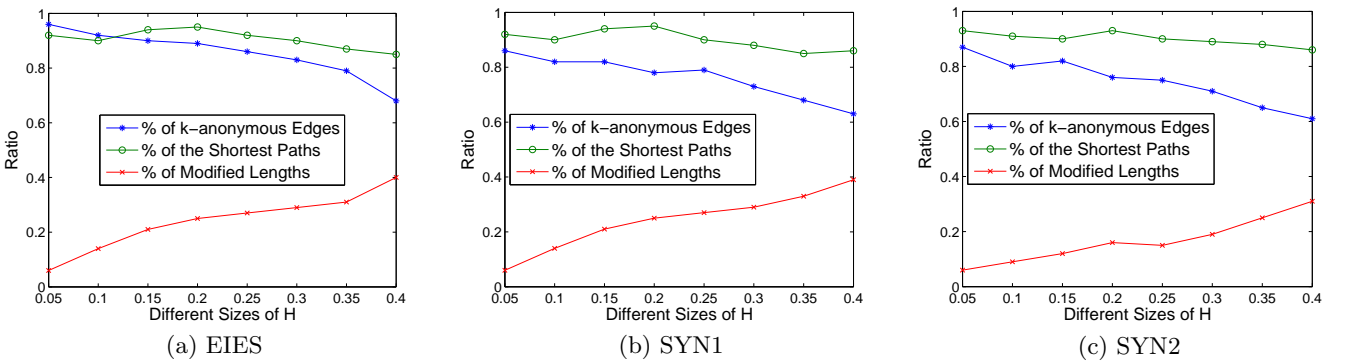


Figure 5: The comparison about the three criteria on three cases in the condition of $k=3$ and $\mu=10$. The blue star line denotes the percentage of k-anonymous edges, the green circle line means the percentage of the preserved shortest paths, and the red marked line stands for the ratio of length differences between the original ones and the modified ones.

ones is more affected by the size of H rather than k .

6. CONCLUSIONS

In consideration of the privacy issue in social network data mining applications, the link's weights between social network entities are sensitive in some cases such as in the business transaction expenses. This paper addresses a balance between the protection of sensitive weights of network links (edges) and two global structure utilities, the shortest paths and the corresponding shortest path lengths.

In this paper, we presented one algorithm based on random walk and matrix analysis to modify individual (sensitive) edge weights and try to keep exactly the same shortest paths as well as their lengths close to those of the original social network. Our experimental results demonstrate that our proposed modification strategy does meet the expectation of our mathematical analysis.

Further research work along this line can be carried out to extend our modification strategies to modify weights of the original edges in case of a particular social network in which the social network structure and its weights change over time.

7. REFERENCES

- [1] A. Acquisti and R. Gross. Privacy risks for mining online social networks. In NSF Symposium on Next Generation of Data Mining and Cyber-Enabled Discovery for Innovation (NGDM 2007), Baltimore, MD, October 2007.
- [2] L. Backstrom, C. Dwork, and J. Kleinberg. Wherefore art thou R3579X? anonymized social networks, hidden patterns, and structural steganography. In Proceedings of the 16th International Conference on World Wide Web, Alberta, Canada, pp. 181-190, 2007.
- [3] F. C. Chang. Inversion of a perturbed matrix. Applied Mathematics Letters, 19: 169-173, 2006.
- [4] G. Cormode, D. Srivastava, T. Yu, and Q. Zhang. Anonymizing bipartite graph data using safe groupings. In Proceedings of VLDB 2008, pp. 833-844, Auckland, New Zealand, Aug 23-28, 2008.
- [5] M. Hay, G. Miklau, D. Jensen, D. F. Towsley, and P. Weis. Resisting structural re-identification in anonymized social networks. In Proceedings of VLDB 2008, pp. 102-114, Auckland, New Zealand, Aug 23-28, 2008.
- [6] M. Hay, G. Miklau, D. Jensen, P. Weis, and S.

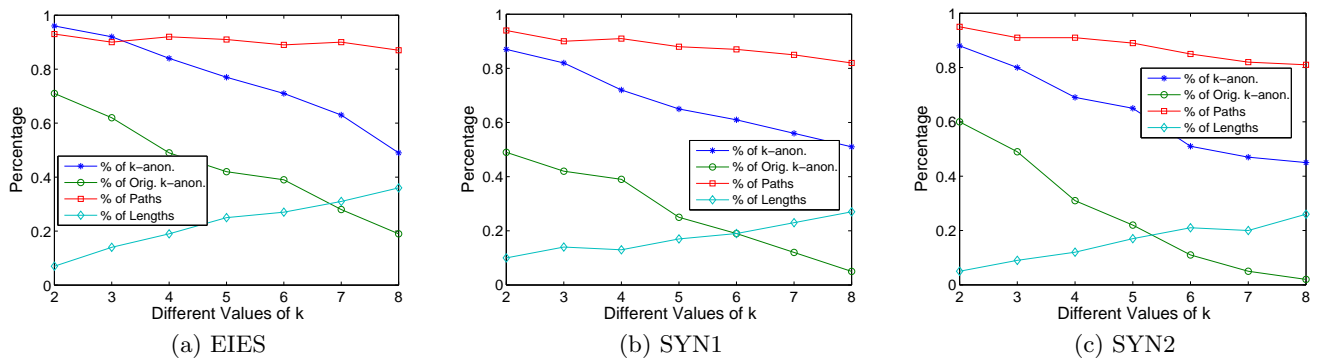


Figure 6: The comparison about the percentage of k -anonymous edges in the condition of $H=10\%$ and $\mu=10$. The blue star line denotes the percentage of k -anonymous edges in modified networks, the green circle line means the percentage of k -anonymous edges in original networks, the red squared line stands for the preserved shortest paths, and the cyan diamond line is the ratio of length differences between the original ones and the modified ones.

Srivastava. Anonymizing social networks. University of Massachusetts, Amherst, MA, Tech. Rep. 07-19, 2007.

[7] T. Heimo, J. Kumpula, K. Kaski, and J. Saramaki. Detecting modules in dense weighted networks with the Potts method. arXiv:0804.3457, 2008.

[8] A. Inkpen. The Japanese corporate network transferred to North America: implications for North American firms. *The International Executive*, 36(4): 411-433, 1994.

[9] A. Korolova, R. Motwani, S. Nabar, and Y. Xu. Link privacy in social networks. In *Proceedings of IEEE 24th International Conference on Data Engineering (ICDE 2008)*, pp. 1355-1357, Cancun, Mexico, Apr 7-12, 2008.

[10] K. Liu, K. Das, T. Grandison, and H. Kargupta. *Privacy-Preserving Data Analysis on Graphs and Social Networks*. In *Next Generation Data Mining*. Chapter 21, pages 419-437. Edited by Hillol Kargupta, Jiawei Han, Philip Yu, Rajeev Motwani, and Vipin Kumar, CRC Press, Dec 2008.

[11] K. Liu and E. Terzi. Towards identity anonymization on graphs. In *Proceedings of SIGMOD 2008*, pp. 93-106, Vancouver, BC, Canada, Jun 9-12, 2008.

[12] L. Liu, J. Wang, J. Liu, and J. Zhang. Privacy preservation social networks with sensitive edge weights. In *Proceedings of the 2009 SIAM International Conference on Data Mining (SDM 2009)*, Sparks, Nevada, April 30–May 2, 2009, to appear.

[13] L. Sweeney. K -anonymity: a model for protecting privacy. *International Journal on Uncertainty, Fuzziness and Knowledge-based Systems*, 10(5):557-570, 2002.

[14] J. Tang, D. Zhang, L. Yao, J. Li, L. Zhang, and Z. Su. ArnetMiner: extraction and mining of academic social networks. In *Proceeding of the 14th ACM SIGKDD*, pp. 990-998, Las Vegas, Nevada, USA, Aug 24-27, 2008.

[15] D. Wang, C. Liau, and T. Hsu. Privacy protection in social network data disclosure based on granular computing. In *Proceedings of the 2006 IEEE International Conference on Fuzzy Systems*, pp. 997-1003, Vancouver, BC, Canada, July 16-21, 2006.

[16] L. Yen, M. Saerens, A. Mantrach, and M. Shimbo. A family of dissimilarity measures between nodes generalizing both the shortest-path and the commute-time distances. In *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD 2008)*, pp. 785-793, Las Vegas, NV, USA, Aug 24-27, 2008.

[17] X. Ying and X. Wu. Randomizing social networks: a spectrum preserving approach. In *Proceedings of the 2008 SIAM International Conference on Data Mining*, pp. 739-750, Atlanta, GA, Apr 24-26, 2008.

[18] E. Zheleva and L. Getoor. Preserving the privacy of sensitive relationships in graph data. In *Proceedings of the 1st ACM SIGKDD International Workshop on Privacy, Security, and Trusting KDD*, San Jose, California, pp. 153-171, Aug 2007.

[19] B. Zhou and J. Pei. Preserving privacy in social networks against neighborhood attacks. In *Proceedings of the 24th International Conference on Data Engineering (ICDE'08)*, Cancun, Mexico, pp. 506-515, April 2008.