TRANSACTIONS ON DATA PRIVACY 7 (2014) 71–97

Privacy Preserving Social Network Publication Against Mutual Friend Attacks

Chongjing Sun*, **Philip S Yu****, **Xiangnan Kong****, **Yan Fu***

*Web Science Center, School of Computer Science and Engineering, University of Electronic Science and Technology of China, Chengdu, 611731, China.

**Department of Computer Science, University of Illinois at Chicago, Chicago, IL 60612, USA.

E-mail: chingsun00@gmail.com, psyu@cs.uic.edu, xkong4@cs.uic.edu, fuyan@uestc.edu.cn

Abstract. Publishing social network data for research purposes has raised serious concerns for individual privacy. There exist many privacy-preserving works that can deal with different attack models. In this paper, we introduce a novel privacy attack model and refer it as a mutual friend attack. In this model, the adversary can re-identify a pair of friends by using their number of mutual friends. To address this issue, we propose a new anonymity concept, called *k*-NMF anonymity, i.e., *k*-anonymity on the number of mutual friends, which ensures that there exist at least *k*-1 other friend pairs in the graph that share the same number of mutual friends. We devise algorithms to achieve the *k*-NMF anonymity while preserving the original vertex set in the sense that we allow the occasional addition but no deletion of vertices. Further we give an algorithm to ensure the *k*-degree anonymity in addition to the *k*-NMF anonymity. The experimental results on real-word datasets demonstrate that our approach can preserve the privacy and utility of social networks effectively against mutual friend attacks.

Keywords. privacy-preserving; social network; data publication; mutual friend.

1 Introduction

With the advance on mobile and Internet technology, more and more information is recorded by social network applications, such as Facebook and Twitter. The relationship information in social networks attracts researchers from different academic fields. As a consequence, more and more social network datasets were published for research purposes [1]. The published social network datasets may incur the privacy invasion of some individuals or groups. With the increasing concerns on the privacy, many works have been proposed for the privacy-preserving social network publication [2, 3].

Tai and Yu proposed the friendship attack model [4], which addressed the issue that an attacker can find out not only the degree of a person, but also the degree of his friend.

*This research work was supported in part by the National Natural Science Foundation of China under Grant No.61003231 and No.61103109, the research funds for central universities under grant No. ZYGX2012J085 and ZYGX2011J057, Huawei university-enterprise cooperation project YBCB2011057 and US NSF through grants CNS-1115234, DBI-0960443, and OISE-1129076.

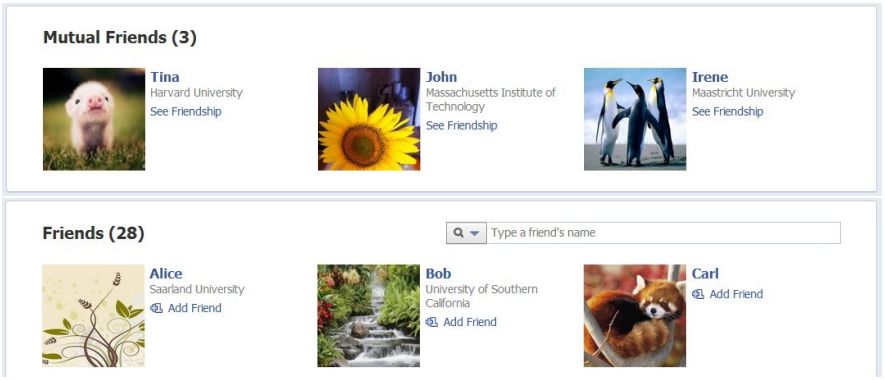


Figure 1: Friend lists on Facebook

It solves the attacks based on the degrees of two connected vertices. But it is not sufficient to just protect against the friendship attack as there are more information available on the social network. For example, the graph in Fig. 3(a) is a k^2-degree anonymized graph with $k = 2$. The number around each vertex represents the degree of this vertex. If an attacker can obtain the number of mutual friends between two connected vertices, he still can identify (D, F) from other friend pairs, as only (D, F) has 2 mutual friends. This will be explained in more details later. In most social networking sites, such as Facebook, Twitter, and LinkedIn, the adversary can easily get the number of mutual friends of two individuals linked by a relationship. As shown in Figure 1, one can directly see mutual friend list shared with one of his friends on Facebook. Usually, the adversary can get the friend lists of two individuals from Facebook, such as the friend list in Figure 1, and then get the number of mutual friends by intersecting their friend lists.

In this paper, we introduce a new relationship attack model based on the number of mutual friends of two connected individuals, and refer it as a *mutual friend attack*. Figure 3 shows an example of the mutual friend attack. The original social network G with vertex identities is shown in Figure 3(b), and can be naively anonymized as the network G' shown in Figure 3(c) by removing all individuals' names. The *number* on each edge in G' represents the number of mutual friends of the two end vertices. Alice and Bob are friends, and their mutual friends are Carl, Dell, Ed and Frank. So the number of mutual friends of Alice and Bob is 4. After obtaining this information, the adversary can uniquely re-identify the edge (D, H) is $(Alice, Bob)$. Also, $(Alice, Carl)$ can be uniquely re-identified in G'. By combining $(Alice, Bob)$ and $(Alice, Carl)$, the adversary can uniquely re-identify individuals Alice, Bob and Carl. This simple example illustrates that it is possible for the adversary to re-identify an edge between two individuals and maybe indeed identify the individuals when he can get the number of mutual friends of individuals. Note that we do not consider the mutual friend number of two nodes if they are not connected. For convenience, we say the *number of mutual friends of two nodes* connected by an edge e as the *number of mutual friends of* e.

In order to commit a mutual friend attack, the adversary only need to acquire the number of mutual friends of two victims. Based on this kind of simple background knowledge, an adversary can issue the attack on published social network to re-identify the edge corresponding to the relationship between two victims as well as associated edge information, such as email content, the weight reflecting sensitive transaction expenses [20]. From the

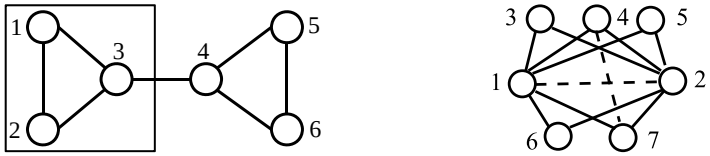


Figure 2: Examples of the k-NMF anonymization

above example, we can see that the adversary may indeed re-identify an individual by combing some mutual friend attacks. Therefore, the sensitive vertex information associated with an individual, such as religious beliefs, hobbies, will be disclosed to the attacker.

To protect the privacy of relationship from the mutual friend attack, we introduce a new privacy-preserving model, k-anonymity on the number of mutual friends (k-NMF Anonymity). For each edge e, there will be at least k-1 other edges with the same number of mutual friends as e. It can be guaranteed that the probability of an edge being identified is not greater than $1/k$. We propose algorithms to achieve the k-NMF anonymity for the original graph while preserving the original vertex set in the sense that we allow the occasional addition but no deletion of vertices. As pointed out in [4, 29], by preserving the original vertex set, various analysis on the anonymized graph, such as identifying vertices providing specific roles like centrality vertex, influential vertex, gateway vertex, outlier vertex, etc., will be more meaningful. The experimental results show that our approaches can maintain much of the utility of social networks evaluated by some commonly used metrics [4, 21–26, 29] concerning the network characteristics, such as the clustering coefficient, average path length, and betweenness centrality.

Challenges. As the k-NMF anonymity model is more complicated than the k-degree anonymity model, more challenges need to be handled. First, adding or removing a different edge may affect a different number of edges on their mutual friends. In the k-degree anonymity model, the adversary attacks using the degree of the vertex. Adding an edge only increase the degrees of the two end vertices of this edge. In the k-NMF anonymity model, the adversary attacks using the number of mutual friends. Adding an edge can increase the numbers of mutual friends of many edges. In Figure 3(b), adding an edge between Dell and Frank will affect the NMFs of (Dell, Alice), (Frank, Alice), (Dell, Bob), (Frank, Bob), and (Dell, Frank). Second, we need to provide a criterion on choosing where to add or delete the edge while considering the utility of the graph. In fact, we map the k-NMF anonymization problem into an edge anonymization problem in contrast to the vertex anonymization problem in the k-degree anonymization. Edges are anonymized one by one. Adding or deleting an edge should not destroy the anonymization of the already anonymized edges. To anonymize an edge, we can get many candidate edge operations and need to choose the best one. Besides, we need to consider the impact of the newly added edges on the number of mutual friends.

Contributions. Our contributions can be summarized as follows. (1) We introduce the k-NMF problem and formulate it as an edge weight anonymization problem where the edge weight is the NMF of the two end vertices. (2) We explore the geometry property of the graph to devise effective anonymization algorithms while preserving the vertex set to achieve better utility. We introduce the principle of preserving anonymized triangles in the graph to avoid the problem of repeatedly re-anonymizing edges during the anonymization process. (3) For the edge addition, we use the breadth-first manner to preserve util-

ity. We also introduce the maximum mutual friend criterion to break the tie on selecting candidate vertex to connect. (4) For the edge deletion, we explore the triangle linking property to delete edges between vertices already belonging to a triangle connection in the network to avoid repeated re-anonymization of edges. (5) We devise an algorithm which can anonymize the *k*-NMF anonymized graph to simultaneously satisfy the *k*-degree anonymity, while preserving the vertex set. (6)The empirical results on real datasets show that our algorithms perform well in anonymizing the real social networks.

The rest of the paper is organized as follows. We firstly introduce the related work in Section 2. Then we define the problem and design algorithms to solve it in Section 3 and 4. We conduct the experiments on real data sets and conclude in Section 5 and 6.

2 Related Work

The data sharing or publishing leads to the privacy protection issues. Early works mainly focus on privacy-preserving data publishing on relational data. The attack models can be divided into three categories, record linkage, attribute linkage, and table linkage. In the attack of record linkage, an adversary try to re-identify the owner of a record based on the quasi-identifers QID. The records with the same QID value forms a group. If the number of records in a group is very small, the attacker could identify the victim's record from this group with high probability. To prevent such kind of attacks, the k-anonymity was proposed by Samarati and Sweeney [5]. In this privacy model, there are at least k records in each qid group. Then the probability of a record be re-identified is not greater than $1/k$. The k-anonymity model cannot provide protection under the attack of attribute linkage. If all the records in a qid group have the same sensitive attribute value v, then the owner of record in this group has value v with probability as 100%. The l-diversity model [6] ensures that there exist l distinct sensitive values in each qid group. But this privacy model does not consider the distribution of sensitive values in a qid group. The t-closeness model [7] requires that the distribution of sensitive attribute in each group is similar to each other among the whole dataset. In the attack of table linkage, an attacker try to make sure whether the record of an individual exist in a released data table, and the corresponding privacy model is δ-presence [8], which limit the probability of the presence be no more than δ. A comprehensive survey on the privacy-preserving data publishing can be found in [9].

In the past few years, the privacy preservation is extended to the social network publication with the blooming research on the social network analysis. The privacy-preserving approaches can be divided into two categories, the clustering-based approaches and graph-editing-based approaches. The clustering-based approaches firstly design method to cluster the vertices into different groups, and then replace each group with a super vertex. The edges between two groups are represented by a super edge. If each super vertex in the clustered graph contains at least k original vertices, the probability of re-identify a user is no more than $1/k$. Hay et al. [10] were the first to put forward the clustering-based approach to defend the attacks based on the vertex refinement, subgraph queries, and hub fingerprints. Further, Campan and Truta [11] designed algorithms to cluster the vertices while considering the information lost on the vertex label and graph structure. Zheleva and Getoor [12] proposed the clustering-based approach to preserve the sensitive link. Cormode et al. [13] and Bhagat et al. [14] put forward the (k, l)-clustering model based on the bipartite graph. Campan et al. [15] devised a p-Sensitive k-Anonymity model to ensure that each cluster contains at least p distinct sensitive values. The clustering-based approaches only publish the clustered graph with super vertices and edges. It significantly reduce the utility of the

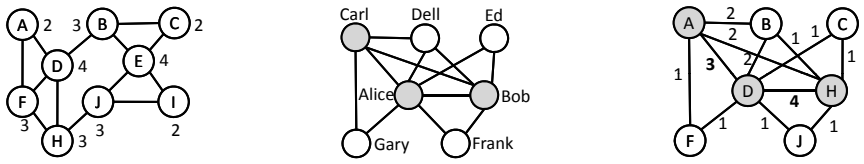


Figure 3: Mutual friend attack in a social network

original network.

The graph-editing-based approaches can be divided into two categories, the randomization-based approaches and k-anonymity-based approaches. Hay et al. [16] proposed a Rand Add/Del method, which randomly deletes k edges from the network and then randomly adds k new edges into the network. In addition to using the Rand Add/Del method, Ying and Wu [17] applied the Rand Switch method to randomize the original graph. It switch the existed edges (u, v) and (t, w) as non-existed edges (t, v) and (u, w). Ying and Wu [17] designed Spctr Rand Add/Del and Spctr Rand Switch methods to maintain most of the spectral characteristics of original networks. Hanhijarvi et al. [18] and Ying et al. [19] applied the Rand Switch method to keep the degree sequence unchanged while preserving the privacy. They constructed the Markov chain based on the edge switched graphs, and used the Metropolis-Hastings sampling to find a randomized graph which can maintain the clustering coefficient or average path length of original networks. These methods cannot provide a quantifiable guarantee on the privacy protection such as the k-anonymity provided.

The k-anonymity-based approaches implement k-anonymity on different background knowledge, and these approaches are the most related to our work. Backstorm et al. [21] pointed out that simply removing identities of vertices cannot guarantee privacy. Many works have been done to prevent the vertex re-identification with the vertex degree. Liu et al. [22] studied the k-degree anonymization which ensures that for any node v there exist at least k-1 other vertices in the published graph with the same degree as v. Tai et al. [4] introduced a friendship attack, in which the adversary uses the degrees of two end vertices of an edge to re-identify victims. Associated with community identity for each vertex, in [29] they proposed the k-structural diversity anonymization, which guarantees the existence of at least k communities containing vertices with the same degree for each vertex. As these works only focus on the vertex degree, they cannot achieve the k-NMF anonymity, which focuses on the number of common neighbors of two vertices.

Many works have also been done to prevent the vertex re-identification based on the subgraph structural information. Zhou and Pei [23] proposed a solution to battle the adversary's 1-neighborhood attacks. Cheng et al. [24] proposed the k-isomorphism model, which disconnects the original graph into k-isomorphic subgraph. To protect against multiple structural attacks, Zou et al. [25] proposed the k-automorphism model, which converts the original network into a k-automorphic network. But it does not prevent the mutual friend attack. The network in Figure 2(a) satisfies the 2-automorphism, but the edge $(3, 4)$ is not protected under the mutual friend attack. This is because the edge $(3, 4)$ does not have mutual friends while all the others have one. Wu et al. [26] proposed the k-symmetry model, which gets a k-automorphic network by orbit copying. All these algorithms need to introduce many new vertices and adjust many edges to achieve their targets. Therefore, the utility of the original graph will be decreased too much. In any case, these works are

aimed at different types of attack model from ours as illustrated in Figure 2(a).

Considering the sensitive labels of vertices, Zhou et al. [28] and Tai et al. [29] proposed the l-diversity model, which ensure that the vertices in each group satisfying k-anonymity contains at least l distinct sensitive labels. Other works focus on the problem of link disclosure [30, 31], which decides whether there exists a link between two individuals. It is different from the relationship re-identification introduced in this section. Besides, the differential privacy is attempted to enforce on the social network data [32–39].

3 Problem definition

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

Definition 1. ***The NMF of an edge.*** For an edge e between two vertices v_1 and v_2 in a graph $G(V, E)$, i.e., $v_1, v_2 \in V$, $e \in E$ and $e = (v_1, v_2)$, the number of mutual friends of the edge e is the number of mutual friends of v_1 and v_2.

Let $\boldsymbol{f}$ be the *number sequence of mutual friends* for G, in which entries are sorted in descending order, i.e., $\boldsymbol{f}_1 \geq \boldsymbol{f}_2 \geq ... \geq \boldsymbol{f}_m$. Let $\boldsymbol{l}$ be the list of edges corresponding to $\boldsymbol{f}$, i.e., $\boldsymbol{f}_i$ is the NMF of the edge $\boldsymbol{l}_i$. For example, in Figure 4(c), $\boldsymbol{f} = \{2, 2, 2, 2, 1, 1, 1, 1\}$, and $\boldsymbol{l} = \{(v_1, v_3), (v_2, v_3), (v_3, v_4), (v_3, v_5), (v_3, v_4), (v_3, v_5), (v_1, v_2), (v_1, v_4), (v_2, v_5), (v_4, v_5)\}$. Similar to the power law distribution of the vertex degree [40], the NMF also has the same property [41].

Property 2. Scale free distribution of NMFs [41]**.** *The NMFs of edges in the large social network often have a scale-free distribution, which means that the distribution follows a power law or at least asymptotically.*

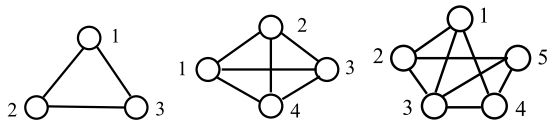
Definition 3. ***Mutual friend attack***. Given a social network $G(V, E)$ and the anonymized network $G'(V', E')$ for publishing. For an edge $e \in E$, the adversary can get the number $\boldsymbol{f}_e$ of mutual friends of e. Mutual Friend Attack will identify all *candidate edges* $e' \in E'$ with the number $\boldsymbol{f}_{e'}$ of mutual friends as $\boldsymbol{f}_e$.

Suppose that the candidate edge set of an edge e is $E'_e = \{e' | e' \in E', \boldsymbol{f}_{e'} = \boldsymbol{f}_e\}$. An adversary re-identifies the edge e with high confidence if the number of candidate edges is too small. Hence, we set a threshold k to make sure that for each edge $e \in E$, the number of candidate edges is no less than k, i.e., $|E'_{\boldsymbol{f}_e}| \geq k$. We define the *k*-anonymous sequence before defining the *k*-NMF anonymous graph.

Definition 4. ***k-anonymous sequence***[22]. A sequence vector $\boldsymbol{f}$ is k-anonymous, if for any entry with value as v, there exist at least $k - 1$ other entries with value as v.

Definition 5. ***k*-NMF**. A graph $G'(V', E')$ is k-NMF anonymous if the number sequence $\boldsymbol{f}'$ of mutual friends of edges in G' is a k-anonymous sequence.

Definition 5 states that for each edge $e \in E$, the number of candidate edges in G' is no less than k. Consider the graphs in Figure 5 as an example. There are three edges in Figure 4(a), and the NMFs of all these edges are equal to 1. Hence, this graph is a 3-NMF anonymous graph. As the six edges in the graph of Figure 4(b) have 2 mutual friends, this graph is a 6-NMF anonymous graph. The graph in Figure 4(c) has four edges



(a) 3-NMF (b) 6-NMF (c) 4-NMF

Figure 4: Examples of k-NMF anonymous graph

$(v_1, v_3), (v_2, v_3), (v_3, v_4), (v_3, v_5)$ with the NMF as 2, and the NMFs of other four edges are equal to 1. Hence, this graph is a 4-NMF anonymous graph. Some properties on the number of mutual friends are described as follows.

Proposition 6. *Given a graph $G(V, E)$, the number of mutual friends of an edge $e \in E$ is equal to the number of triangles containing e in G .*

Take the graph in Figure 4(c) as an example. The mutual friends of vertices v_2 and v_3 are v_1 and v_5, so the number of mutual friends of the edge $e = (v_2, v_3)$ is 2. It is equal to the number of triangles containing e. These triangles are (v_1, v_2, v_3) and (v_2, v_3, v_5).

Proposition 7. *Let $G(V, E)$ be a graph and f be the number sequence of mutual friends of edges in G, where $|E| = m$. Then $\sum_{i=1}^{m} \boldsymbol{f}_i = 3n_{\triangle}$, where $n_{\triangle}$ is the number of triangles in G and $\boldsymbol{f}_i$ is the number of mutual friends of the i-th edge.*

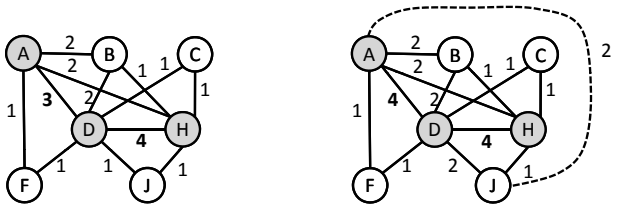
Different from the degree sequence in previous work [22], which can maintain the number of entries in the sequence, the number sequence of mutual friends will have more entries added into it when new edges are added into the graph. Besides, according to Propositions 6 and 7, the number of mutual friends is related to the number of triangles in the graph. Therefore, adding one edge will affect the NMF of many edges, and adding a different edge may affect the NMF of a different number of edges. This can be illustrated by an example shown in Figure 2(b). After we add the edge $(1, 2)$, the NMFs of all ten edges increase by one. If we add the edge $(4, 7)$, only the NMFs of edges $(1, 4), (1, 7), (2, 4)$, and $(2, 7)$ increase by one. Therefore, one cannot anonymize a graph by simply minimizing the number of changed edges.

Anonymized Triangle Preservation Principle (ATPP). In our algorithms, we anonymize the edges in the graph one by one. An *anonymized triangle* is a triangle with some edges already anonymized in the process of the graph anonymizing. The *Anonymized Triangle Preservation Principle* aims to preserve the anonymized triangles containing already anonymized edges. It means that we neither create some additional anonymized triangles via edge addition nor destroy any via edge deletion.

Creating (destroying) a triangle containing an already anonymized edge by edge addition (deletion) will increase (decrease) the NMF of this edge, indeed destroy the anonymization of this edge. This leads to repeatedly anonymization of this edge. By preserving the anonymized triangles, we can avoid this problem during the anonymization process.

Definition 8. *k*-NMF anonymization problem. Given a graph $G(V, E)$ and an integer k, the problem is to anonymize the graph G to a k-NMF anonymous graph G' with edge addition and deletion, such that the vertex set of the original graph G is preserved.

For example, in the original social network G in Figure 5(a), both edges (A, D) and (D, H) can be uniquely identified by their NMFs. Hence, G does not satisfy the 2-NMF anonymous



(a) Original network G (b) 2-NMF anonymous G'

Figure 5: An example of k-NMF anonymization

condition. In order to make the original social network G satisfying the 2-NMF anonymous, we can have many edge operations. In these operations, we can add an edge between vertices A and J, and transform G into a 2-NMF anonymous network G' in Figure 5(b). Therefore, only by an edge addition, we can finish the 2-NMF anonymization problem. The k-NMF anonymization problem will try to find a solution which can transform the original social network G into a new network G' satisfying the k-NMF anonymous condition and change the original social network G as little as possible.

4 *k*-NMF anonymization approach

In the above section, we have shown that changing one edge may affect the NMFs of other edges. To handle this challenge, we utilize the scala free distribution property shown in Property 2, and introduce the principle of preserving the anonymized triangles. By exploring the geometry property of the graph, we devise two effective anonymization algorithms to preserve the utility while satisfying the k-NMF anonymity.

4.1 Algorithm ADD

In this subsection, we aim to anonymize the original graph only by edge addition. We organize edges into groups, and anonymize the edges in the same group to have the same NMF. The k-anonymity requires there exist at least k edges in a group. The main process of k-NMF anonymization is shown in Figure 6. First, more than k edges are put into an edge group, and each edge of this group is anonymized by the proposed BFSEA method in the following section. Then the edges in the social network are gradually put into some edge group and anonymized by the BFSEA. The edges in the edge list are dynamically updated as some new edges are added into the network while some are removed. Hence, the k-NMF anonymization mainly focuses on two problems, "how to group edges" and "how to anonymize edges in a group". Property 2 states that the NMFs of edges in large social networks follow a scala free distribution. Hence, only a small number of edges have a high NMF. We first anonymize these edges, and many edges with low NMF do not need to be processed. Next, we consider how to group edges.

4.1.1 Group edges

Suppose the original graph is $G(V, E)$ and the gradually anonymized graph is $G'(V', E')$. Initially, we sort the NMF sequence $\boldsymbol{f}$ in descending order and construct the corresponding edge list $\boldsymbol{l}$ as described in Section 3. We mark all edges as "unanonymized", and then

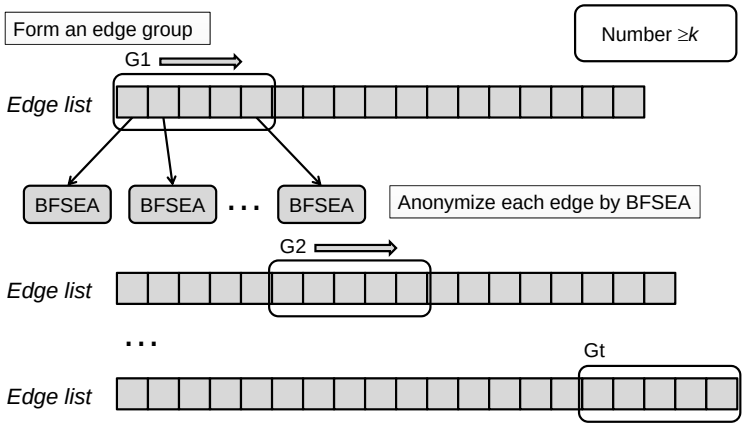


Figure 6: The illustration of k-NMF anonymization algorithms

anonymize the edges one by one. Iteratively, we start a new group GP with the group NMF, g_f, equal to the NMF of the first unanonymized edge in $\boldsymbol{l}$. Then we select the edges with NMF equal to g_f and mark them as "anonymized". We iteratively select the first unanonymized edge in $\boldsymbol{l}$ and anonymize it by adding edges to increase its NMF to g_f. After anonymizing this edge, we mark it as "anonymized" and put it into GP. Adding new edges affects the NMF of some other edges, and these new edges will be added into $\boldsymbol{f}$ and $\boldsymbol{l}$. Hence we resort the sequences $\boldsymbol{f}$ and $\boldsymbol{L}$ after each edge is anonymized. Algorithm 1 shows the detailed description of the ADD algorithm. Next, we consider when we start another new group.

An intuitive method, named **IntuitGroup**, starts another group when the number of edges in the group GP is equal to k. Alternatively, to consider the anonymization cost, we propose a greedy method to decide when we start another group after $|GP| \geq k$, named **GreedyGroup**. Suppose that $\boldsymbol{f}^{(u)} \subseteq \boldsymbol{f}$ is the NMF sequence corresponding to the unanonymized edge list $\boldsymbol{l}^{(u)} \subseteq \boldsymbol{l}$. Notice that $\boldsymbol{f}^{(u)}$ and $\boldsymbol{l}^{(u)}$ are dynamically updated with $\boldsymbol{f}$ and $\boldsymbol{l}$ after anonymizing each edge. Similar to the consideration in [22], after putting k edges into GP, GreedyGroup iteratively checks whether it should merge the edge $\boldsymbol{l}_1^{(u)}$ into GP or start another group. The decision is made according to the following two costs based on the number of added mutual friends.

$$C_{merge} = (g_f - \boldsymbol{f}_1^{(u)}) + I(\boldsymbol{f}_2^{(u)}, \boldsymbol{f}_{k+1}^{(u)}) \tag{1}$$

$$C_{new} = I(\boldsymbol{f}_1^{(u)}, \boldsymbol{f}_k^{(u)}) \tag{2}$$

where $I(\boldsymbol{f}_i^{(u)}, \boldsymbol{f}_j^{(u)}) = \sum_{l=i}^{j}(\boldsymbol{f}_i^{(u)} - \boldsymbol{f}_l^{(u)})$.

For Eq.(1), we put $\boldsymbol{l}_1^{(u)}$ into GP. $\boldsymbol{l}_1^{(u)}$ has $\boldsymbol{f}_1^{(u)}$ mutual friends, so we need to add $g_f - \boldsymbol{f}_1^{(u)}$ mutual friends for anonymizing $\boldsymbol{l}_1^{(u)}$. To satisfy k-anonymity, we need to put at least k edges into a new group GP'. Hence we put edges $\boldsymbol{l}_2^{(u)}, ..., \boldsymbol{l}_{k+1}^{(u)}$ into GP'. As we only adding edges, the group NMF of GP' is the maximum NMF among $\boldsymbol{f}_2^{(u)}, ..., \boldsymbol{f}_{k+1}^{(u)}$, i.e., $\boldsymbol{f}_2^{(u)}$. To anonymize $\boldsymbol{l}_i^{(u)}$, $\boldsymbol{f}_2^{(u)} - \boldsymbol{f}_i^{(u)}$ mutual friends need to be added. For Eq.(2), we put $\boldsymbol{l}_1^{(u)}, ..., \boldsymbol{l}_k^{(u)}$ into a new group GP', and the group NMF of GP is $\boldsymbol{f}_1^{(u)}$. Hence C_{merge} is the cost for anonymizing $k+1$ edges while C_{new} is for k edges. So if C_{merge} is less than C_{new}, we anonymize $L_1^{(u)}$

Algorithm 1 The ADD Algorithm (GreedyGroup)

Input: Original graph $G(V, E)$, k
Output: k-NMF anonymized graph $G'(V', E')$
Initialization: $G' = G$, and mark all edges as "unanonymized". Compute and sort the sequences $\boldsymbol{f}$ and $\boldsymbol{l}$. $\boldsymbol{f}^{(u)} = \boldsymbol{f}, \boldsymbol{l}^{(u)} = \boldsymbol{l}, G_f = \emptyset$

```
while l^(u) ≠ ∅ do
  if |l^(u)| < 2k then do cleanup-operation and break.
  GP = {e|e ∈ l^(u) and f_e^(u) = f_1^(u)}; g_f = f_1^(u); G_f = G_f ∪ g_f.
  Mark any e ∈ GP as "anonymized"; update f^(u) and l^(u).
  while |GP| < k or (|GP| ≥ k and C_merge ≤ C_new) do
    Anoymize l_1^(u) by BFSEA. GP = GP ∪ l_1^(u), update l^(u) and f^(u).
  end while
end while
return G'(V', E').
```

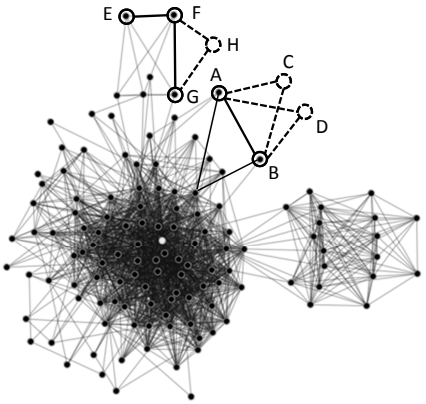


Figure 7: An example of cleanup operation

and merge it into GP, and check the next unanonymized edge. Otherwise we start another new group with $\boldsymbol{l}_1^{(u)}$.

For each edge e, GreedyGroup looks ahead at $O(k)$ other edges to decide whether merging e with this group or starting a new group. Therefore, the time complexity of GreedyGroup is $O(k|E|)$.

4.1.2 Cleanup-operation

In each iteration of the ADD algorithm, it checks the number of unanonymized edges, n_u. If $n_u < 2k$, the remaining edges are put into a group; and if $n_u < k$, $k - n_u$ edges needed to be added following the ATPP, so these k edges can form a group. New vertices will be added into the graph if the ATPP cannot be satisfied. Finally, we have an unanonymized edge set E_u, and need to anonymize them to have the same NMF.

Next, we anonymize the edges E_u in this group. Usually, we set the group NMF as the largest NMF among unanoymized edges, denoted as g_f. Then we sum the difference as $sd = \sum_{e \in E_u}(g_f - \boldsymbol{f}_e)$, where $\boldsymbol{f}_e$ is the number of mutual friends of the edge e, and $g_f = max_{e \in E_u} f_e$. If $sd >= k/2$, then we add sd nodes and $2 \cdot sd$ edges into the graph. That means

that for each unanonymized edge e, we add $g_f - \boldsymbol{f}_e$ vertices and link them with the two end vertices of e. As all the newly added $(2 \cdot sd >= k)$ edges have only one mutual friend, they can form a new group. Then we mark the new edges as "anonymized" and achieve the task. If $sd < k/2$, then we enlarge the group NMF $g_f = g_f + 1$, and repeat the above process.

Take the social network in Figure 7 as an example. We suppose that k is set as 3, and (A, B), (E, F) and (F, G) are the remaining unanonymized edges. Then we put these edges E_u into a group, and g_f is the NMF of edge (E, F), i.e., 3. As $nmf(E, F) = 3$, the edge (E, F) can be marked as "anonymized" directly. The NMF of edge (F, G) is 2, hence we add $g_f - f_{(F,G)} = 1$ vertex H into the network, and link H with both F and G. Similarly, for the edge (A, B), we need to add 2 vertices and link them with A and B as Figure 7. Then all the edges (A, B), (E, F) and (F, G) have 3 mutual friends. Beside, as all the new edges (F, H),(G, H),(A, C), (A, D), (B, C) and (B, D) have only one mutual friend, they can be put into a new group. As $k = 3$, these edges can be marked as "anonymized".

4.2 BFS-based Edge Anonymization(BFSEA)

In this section, we consider how to anonymize an edge by edge addition while preserving the utility. There are three challenges to increase the NMF of an edge via adding edges. First, the added edge should not affect the NMF of already anonymized edges. Secondly, the added edge should minimize the effect on the utility of the graph. Thirdly, the NMF of the newly added edges should not disrupt the current anonymization process which is progressing in descending order of the NMF value.

Before anonymizing an edge (u, v), the ADD algorithm has created some anonymized groups and got a set G_f containing the group NMFs of these groups. Let g_f be the NMF of the current group GP, and we put g_f into G_f. Anonymizing the edge (u, v) means that we need to increase the NMF of (u, v) to the current group NMF g_f, i.e. we need to create some new triangles containing this edge. Then we try to find some candidate vertices and add new edges to create new triangles. Considering the utility of the graph, we find the candidate vertices based on the Breadth First Search (BFS).

From the nodes u and v, *BFS-based Edge Anonymization* traverses the graph in a breadth-first manner. For the i-hop neighbors of u and v, represented by $neig_i(u)$ and $neig_i(v)$, *edge anonymization* finds the candidate vertices from $neig_i(u) \cup neig_i(v)$ and iteratively link the best one with u or v to create a new triangle. We formulize the NMF of the edge (u, v) as $nmf(u, v)$.

4.2.1 Candidates generation

We search the candidate vertices for edge (u, v) in a BFS manner. In the i-hop neighbors of u and v, many vertices cannot be the candidate vertices as violating the ATPP. The vertices w need to satisfy the following conditions to be the candidates in the set CV_i.

a. $w \in neig_i(u) \cup neig_i(v)$.

b. $(w, u, v) \neq \triangle$.

c. $\forall x \in \{u, v\}$ *and* $z \in V'$, if $(w, x) \notin E', (w, z) \in E'$ *and* $(x, z) \in E'$, then (x, z) and (w, z) must be unanonymized.

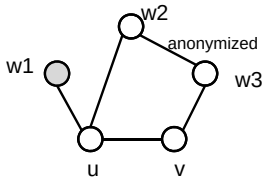


Figure 8: An example of candidates generation

Condition b) states that (u, v, w) is not a complete triangle, which needs to add edges to create a new triangle. This mainly focus on the case when $i = 1$, where w may links with both u and v. Condition c) follows the ATPP, which guarantees that there will be no effect on the already anonymized edges.

We suppose that the graph in Figure 8 is a small part of a social network. During the network anonymiztion, the edge (w_2, w_3) is anonymized. To anonymize the edge (u, v), some new triangles need to be created. As the edge (w_2, w_3) is anonymized, the vertex w_2 cannot be the candidate vertex. This is because if we link the vertices w_2 and v, the anonymization of edge (w_2, w_3) will be destroyed. Hence, only the vertex w_1 can be the candidate vertex for the edge (u, v).

4.2.2 Candidates selection

After getting all the candidate vertices satisfying the conditions, we can add new edges between u,v and $w \in CV_i$ to increase the NMF of (u, v). We iteratively select a vertex from CV_i to increase the NMF of (u, v) until $nmf(u, v)$ reaches g_f or CV_i is empty. If $nmf(u, v) = g_f$, this edge is anonymized successfully.

In each iteration, we need to select the best one which can preserve the most utility of the graph. Based on the link prediction theory [42], we select the candidate vertex w_{max} which guarantees that $nmf'(w_{max}, u) + nmf'(w_{max}, v)$ is maximum, where $nmf'(w, x)$ is defined in Eq.3.

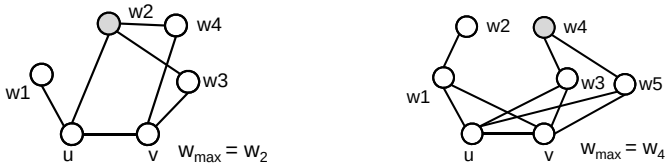
$$nmf'(w, x) = \begin{cases} 0 & (x, w) \in E' \\ nmf(w, x) & otherwise \end{cases} \tag{3}$$

Where $x \in \{u, v\}$. This is referred to as the *maximum mutual friend criterion* for adding edges. The more mutual friends between the two vertices, the less impact the edge addition will have on the utility of the graph.

Figure 9 shows two examples of candidates selection. In Figure 9(a), we try to select the best one from the candidate vertices in the $neig_1(u) \cup neig_1(v)$, i.e., from w_1, w_2, w_3 and w_4. As the number of mutual friends of w_2 and v is maximum, the best candidate for edge (u, v) in G_1 is w_2. The Figure 9(a) shows the example of candidates selection in the 1-hop neighbors while Figure 9(b) in the 2-hop neighbors. In Figure 9(b), the candidate vertices of edge (u, v) are w_2 and w_4. As $nmf(u, w_2) + nmf(v, w_2) = 1 + 0 = 1$ and $nmf(u, w_4) + nmf(v, w_4) = 1 + 2 = 3$, the best candidate for edge (u, v) in Figure 9(b) is w_4.

The selection criteria described in the Eq.3 only can be used for the candidates in the 1-hop and 2-hop neighbors. For all the candidates w in the i-hop neighbors with $i \geq 3$, the NMF of (x, w) is 0. In this situation, we randomly select a candidate vertex w_{max} from CV_i.

As we anonymize edges in descending order of NMF, we must consider the different situations on the NMF of the new edge (x, w_{max}). In the situation $nmf(x, w_{max}) >= g_f$, if $nmf(x, w_{max})$ is not equal to any $g_f^{'} \in G_f$, (x, w_{max}) cannot be added into the graph. This



(a) A simple network G_1 (b) A simple network G_2

Figure 9: Examples of candidates selection

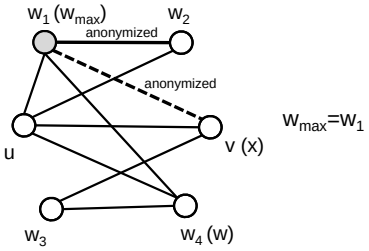


Figure 10: An example of candidate dynamic removal

is because we cannot anonymize this edge in descending order anonymization. Otherwise, we add (x, w_{max}) and mark it as "anonymized". We put this edge into the group with NMF equal to $g_f^{'}$. If $nmf(x, w_{max}) < g_f$, add (x, w_{max}) and mark it as "unanonymized".

4.2.3 Candidates dynamic removal

After a new triangle was created with the vertex $w_{max} \in CV_i$, we need to consider the effect of this triangle on the other candidate vertices in CV_i. To ensure the linking u or v with vertices in CV_i follows the ATPP, some vertices will be dynamically removed from CV_i.

If $w \in CV_i$ connected with w_{max} and the edge (w, w_{max}) is anonymized, then we remove w from CV_i. This is because adding either (w, u) or (w, v) creates a new triangle containing (w, w_{max}), and destroys the anonymization of (w, w_{max}).

For any vertex $w \in CV_i$ with (w, w_{max}) is unanonymized, if (w_{max}, x) is anonymized and $(w, x) \notin E'$, here $x \in \{u, v\}$, then we remove w from CV_i. This is because if we select this w as a new maximum vertex, we need to add (w, x) to create a new triangle containing (u, v), meanwhile we created a triangle containing (w_{max}, x). This destroyed the anonymization of (w_{max}, x).

Take the simple network in Figure 10 as an example. To anonymize the edge (u, v), there are 4 candidate vertices w_1, w_2, w_3 and w_4, and they are put into the candidate set CV_1. Supposed that w_1 is the best candidate w_{max}, and the edge (w_1, w_2) is anonymized. Then after linking the vertices w_1 and v, the vertex w_2 should be removed from the set CV_1. This is because if we link w_2 and v, the anonymiztion of edge (w_1, w_2) will be destroyed, resulting to repeatedly anonymizing edge (w_1, w_2). As the situation in the last part of Section 4.2.2, we suppose that the new edge (w_1, v) is marked as "anonymized". Then we need to remove vertex w_4 from CV_1. This is because if we select w_4 as the next best candidate, we need to link the vertices w_4 and v, which creates a new triangle containing the already anonymized edge w_1, v; this destroys the anonymization of edge (w_1, v).

4.2.4 Edge anonymization

From the nodes u and v, *BFS-based Edge Anonymization* traverses the graph in a breadth-first manner. The *BFSEA* iteratively generates a candidate set CV_i from the i-hop neighbors of u and v, where i increases from 1 to ∞. After getting the candidate set CV_i, *BFSEA* iteratively selects the best one from CV_i by *candidates selection* and creates a triangle to increase the NMF of (u, v), then updates the CV_i by the *candidates dynamic removal*. These operations will break when the NMF of (u, v) reaches the current group NMF g_f or no more candidate vertex can be found from the whole graph.

If $nmf(u, v)$ reaches the current g_f, i.e. (u, v) is anonymized successfully, we mark it as "anonymized". If the *BFSEA* cannot successfully anonymize this edge, adding new vertices can achieve the task. Linking one new vertex with the end vertices of this edge can increase the NMF of this edge by 1. The newly added edges have only one mutual friend, and will be anonymized at the last step of the anonymization algorithm. The above scenario is a pathological case that rarely occurs as in our experiments, no new vertices were added in all cases.

By the breadth-first manner, the *BFSEA* first link u or v with w from the 1-hop neighbors. Thus after (x, w) is added, the shortest path length (SPL) between $x \in \{u, v\}$ and w will only decrease to 1 from 2 with little effect to the utility. Then we gradually increase the value of i, and link u and v with w from the i-hop neighbors, which decreases the SPL between x and w from i to 1. Hence, we prefer the candidates from i-hop neighbors with smaller i value, i.e. breadth-first manner, which can have less effect to the utility.

To get the $neig_i(u)$ and $neig_i(v)$ for every i, we execute the *Breadth-First Search* with the time complexity as $O(|V| + |E|)$. When $i = 1$, we need to compute the $neig_i(u) \cap neig_i(v)$ to ensure $(w, u, v) \neq \triangle$ stated in the *candidates generation*, and the time complexity is $O(|V|)$. When $i \leq 2$, to get the best candidate from CV_i, we compute the $nmf(w, x)$, $x \in \{u, v\}$, with the time complexity as $O(|V|)$. Hence, for each candidate set, the total running time of the NMF computation is $O(|V|^2)$. When $i \geq 3$, we randomly select a candidate from CV_i to create a triangle, and the time complexity is $O(1)$. Hence, the time complexity of *candidates selection* is $O(|V|^2)$. Therefore, the time complexity of *BFSEA* is $O(|V|^2)$.

As there are $O(|E|)$ edges need to be anonymized, the time complexity of the ADD algorithm is $O(|E||V|^2)$.

4.3 Algorithm ADD&DEL

In this subsection, we consider the graph anonymization by edge addition and deletion. Usually, anonymization combining edge deletion with addition will remove or add fewer edges than only applying edge addition. Indeed, it can improve the utility of the anonymized graph. Before introducing the ADD&DEL algorithm, we discuss the method on anonymizing an edge by edge deletion.

Edge-deletion. For an unanonymized edge (u, v), the algorithm finds any candidate edge (x, w), where x is u or v, which satisfies the following conditions.

a. Both (u, w) and (v, w) exist and are unanonymized.

b. For any vertex z linked with x and w, edges (x, z) and (w, z) are still unanonymized.

c. If both (u, w) and (v, w) satisfy condition b), we choose the one with fewer mutual friends.

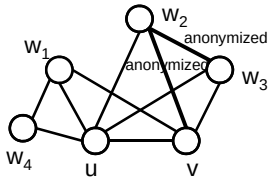


Figure 11: An example of edge-deletion

Condition c) is the reverse of the maximum mutual friend criterion for adding edge. The fewer the mutual friends, the weaker the relationship. Hence dropping the edge has less impact to the utility. After (x, w) is deleted, the shortest path length between x and w will only increase to 2 from 1 with little effect to the utility. Condition a) and b) follows the anonymized triangle preservation principle to guarantee that there will be no effect on the already anonymized edges.

Take the simple network G in Figure 11 as an example.In this network, we suppose that the edges (w_2, w_3) and (w_2, v) are already anonymized. For the edge (u, v), the edges (u, w_2), (u, w_3) and (v, w_3) cannot be the candidate edges. These edges do not satisfy the condition a) or b). For the edges (u, w_1) and (v, w_1), we choose (w_1, v) as the candidate edge according to the condition c).

For an unanonymized edge (u, v), *edge-deletion* initially finds all candidate edges satisfying the edge-deletion conditions, and then puts them into the set CE. During each iteration, the edge $e_{min} \in CE$ with the least mutual friends will be removed from the graph and the set CE. The algorithm stops when the NMF of (u, v) reaches the group NMF g_f or CE becomes an empty set. If CE is empty and the NMF of (u, v) is not equal to g_f, the anonymization of (u, v) is unsuccessful; Otherwise, we mark (u, v) as "anonymized".

The *edge-deletion* is the reverse of the methods in ADD algorithms. The running time mainly costs on the computing of mutual friends, so the complexity of *edge-deletion* is $O(|V|^2)$.

The ADD&DEL Algorithm. This algorithm is shown in Algorithm 2, which anonymizes the graph by edge addition and deletion. Similar to the ADD algorithm, ADD&DEL checks the number of unanonymized edges with NMF equal to the NMF of the first unanonymized edge in sorted sequence $L^{(u)}$. If there are more than k edges, we put them into this group and start another group. Otherwise, we need to anonymize edges to form this group. To gradually anonymize edges and create this group, we initially set the group NMF, g_f, as the mean value of NMFs of the first k unanonymized edges. We record *all initial information* before anonymizing this group. For the unanonymized edge with NMF greater than g_f, we use edge-deletion to anonymize it. If we cannot successfully anonymize this edge, we set $g_f = g_f + 1$ and roll back to *all initial information*. For the unanonymized edge with NMF less than g_f, we apply the ADD algorithm to anonymize it. We gradually anonymize unanonymized edges in sorted sequence $L^{(u)}$ until this group has k edges, and start another group.

In the ADD&DEL algorithm, an edge will be anonymized by either Edge-deletion or methods of the ADD algorithm. Therefore, the time complexity of anonymizing an edge is $O(|V|^2)$, and the time complexity of the ADD&DEL algorithm is $O(|E||V|^2)$.

Algorithm 2 The ADD&DEL Algorithm

Input: Original graph $G(V, E)$, k
Output: k-NMF anonymized graph $G'(V', E')$
Initialization: $G' = G$, and mark all edges as "unanonymized". Compute and sort the sequences $\boldsymbol{f}$ and $\boldsymbol{l}$. $\boldsymbol{f}^{(u)} = \boldsymbol{f}, \boldsymbol{l}^{(u)} = \boldsymbol{l}, G_f = \emptyset$

1: **while** $\boldsymbol{l}^{(u)} \neq \emptyset$ **do**
2: **if** $|\boldsymbol{l}^{(u)}| < 2k$ **then** do *cleanup-operation* and **break**.
3: $EE = \{e | e \in \boldsymbol{l}^{(u)}$ *and* $\boldsymbol{f}_e^{(u)} = \boldsymbol{f}_1^{(u)}\}$;
4: **if** $|EE| \geq k$, **then** new group $GP = EE$, and mark any $e \in GP$ as anonymized, $G_f = G_f \cup \boldsymbol{f}_1^{(u)}$, update $\boldsymbol{l}^{(u)}$ and $\boldsymbol{f}^{(u)}$ and **continue**.
5: $GP = \emptyset$, $g_f = round(mean(\boldsymbol{f}_1^{(u)}, ..., \boldsymbol{f}_k^{(u)}))$. Record all *initial info*.
6: **while** $f_1^{(u)} \geq g_f$ **do**
7: Anonymize $\boldsymbol{l}_1^{(u)}$ by *edge-deletion*.
8: **if** anonymize failed, **then** roll back to *initial info*, and $g_f = g_f + 1$; **else** mark $\boldsymbol{l}_1^{(u)}$ as anonymized and $GP = GP \cup \boldsymbol{l}_1^{(u)}$; update $\boldsymbol{f}^{(u)}$ and $\boldsymbol{l}^{(u)}$.
9: **end while**
10: $G_f = G_f \cup g_f$.
11: **while** $|GP| < k$ **do**
12: Anoymize $\boldsymbol{l}_1^{(u)}$ by *BFSEA*. $GP = GP \cup \boldsymbol{l}_1^{(u)}$, update $\boldsymbol{l}^{(u)}$ and $\boldsymbol{f}^{(u)}$.
13: **end while**
14: **end while**
15: **return** $G'(V', E')$.

4.4 k_1-degree Anonymization Based on k_2-NMF Anonymization

In this subsection, we propose the KDA algorithm on anonymizing the k_2-NMF anonymized graph G' to satisfy k_1-degree anonymity. To maintain the k_2-NMF anonymity of G', the KDA algorithm does not change the NMF of edges in G' when performing anonymization. Proposition 6 stated that the NMF of an edge is related on the number of triangles in which this edge participate, so we anonymize the graph G' without adding new triangles, i.e., the anonymized triangle preservation principle. We can connect two vertices with shortest path length (SPL) no less than three to guarantee that no new triangles will be introduced. Then the NMF of newly added edge is zero. As the degree distribution of the social network follows the power law [40], we only need to anonymize these vertices with large degrees.

The KDA algorithm is similar to the ADD algorithm. The unanonymized vertices are sorted in descending order of their degrees. We gradually group and anonymize them only by edge addition. The vertices in the same group have same degree. To start a new group, KDA set the group degree g_d as the greatest degree of unanonymized vertices. If there are less than k vertices in this group, we anonymize the unanonymized vertices in descending order of their degrees. If this group has more than k vertices, we compute the C_{merge} and C_{new} for the next unanonymized vertex, and decide whether put it into this group or start a new group.

Suppose that the i-hop neighbors of vertex u is $neig_i(u)$. To anonymize the unanonymized vertex u, KDA iteratively and randomly select an unanonymized vertex w_{max} from $neig_3(u)$ and connect u and w_{max}. If the vertex u cannot be anonymized, KDA update the $neig_3(u)$ based on the newly added edges and repeat the above process. If u still cannot be anonymized, we select the candidate vertex from $neig_4(u), neig_5(u)$ and so on until u is anonymized.

When anonymizing a vertex, the KDA algorithm searches the graph in a breadth-first manner to get the candidate vertices. In the worst case, the KDA searches the whole graph and the time complexity is $O(|E| + |V|)$. As there are $O(V)$ vertices needed to be anonymized, the time complexity of the KDA algorithm is $O(|E||V| + |V|^2)$ in the worst case.

5 Experimental Results

In this section, we conduct experiments on real data sets to evaluate the performance of the proposed graph anonymization algorithms. We evaluate the utility of the anonymized graphs by computing their clustering coefficients, average path lengths and betweenness centrality, etc.

5.1 Datasets

We conduct our experiments on four datasets: ACM, Cora, Brightkite, and Gowalla. All datasets are preprocessed into simple undirected graphs without self-loop and multiple edges.

ACM: This dataset was extracted from ACM digital library. We extracted papers published in 12 conference proceedings on computer science before the year 2011. These conferences are ACM Multimedia, OSDI, GECCO, POPL, PODS, PODC, ICCAD, ICSE, ICS, ISCA, ISSAC and PLDI. We derive a graph describing the citations between papers. If one paper cites another paper, an undirected edge will connect both corresponding vertices. The graph includes 7,315 vertices and 16,203 edges after removing the isolated vertices.

Cora: This dataset is composed of a number of scientific papers on computer science [43]. We extract the collaborations between authors to derive the graph. If two authors had co-authored some papers they would be connected. After we removed the authors without any collaboration, the graph contains 14,076 vertices and 72,871 edges.

Brightkite: This dataset shows the friendships between users in the social network Brightkite over the period of April 2008 to October 2010. The graph consists of 58,228 nodes and 214,078 edges, and is available at the SNAP [1].

Gowalla: This dataset contains a friendship network collected from a location-based social networking website. The graph consists of 196,591 nodes and 950,327 edges, and is available at the SNAP [1].

5.2 Evaluation Metrics

In order to measure how well our methods preserving the social network utility, we adopt some common metrics used in [4, 22–25, 29]. Assume we have an original social network $G(V, E)$, and the anonymized social network $G'(V', E')$. We have the following evaluation metrics:

- Average Clustering Coefficient(ACC): evaluates the degree of vertices in a graph tending to cluster together. It is calculated as the average of local clustering coefficients of all the vertices:

$$ACC_{G(V,E)} = \frac{1}{|V|} \sum_{v \in V} C_v \tag{4}$$

where the local clustering coefficient of the vertex v_i is calculated as follows:

$$C_i = \frac{2|\{e_{jk} : v_j, v_k \in N_i, e_{jk} \in E\}|}{k_i(k_i - 1)} \tag{5}$$

where N_i represents the neighbors of vertex v_i, and k_i is the degree of vertex v_i.

- Average Path Length(APL): evaluates the efficiency of information or mass transport on a network. It is defined as the average of shortest path length of all the vertex pairs:

$$APL_{G(V,E)} = \frac{1}{|V|(|V| - 1)} \sum_{u,v \in V, u \neq v} d(u, v) \tag{6}$$

where $d(u, v)$ is the shortest path length of vertices u and v.

- Betweenness Centrality(BC): evaluates the load or importance of a vertex in a network. The betweenness centrality of a vertex v can be calculated as follows:

$$g(v) = \sum_{s \neq v \neq t} \frac{\sigma_{st}(v)}{\sigma_{st}} \tag{7}$$

where σ_{st} is the number of shortest paths from vertex s to vertex t and $\sigma_{st}(v)$ is the number of those paths that pass through vertex v. In this paper, we normalize the betweenness centrality as follows:

$$norm(g(v)) = \frac{g(v) - min(g)}{max(g) - min(g)} \tag{8}$$

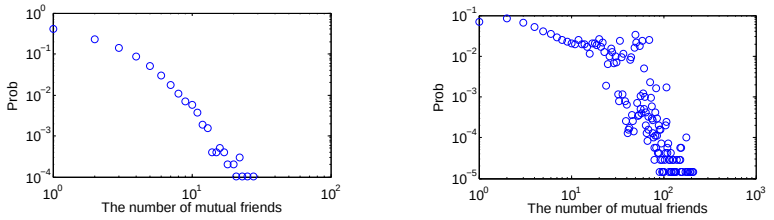
- Percentage of Changed Edges(PCE): evaluates the change on the number of edges in the network. It includes the percentage of add edges (PAE) and percentage of removed edges (PRE). These metrics are computed as follows:

$$PCE_{G,G'} = \frac{|E - E'| + |E' - E|}{|E|} \tag{9}$$

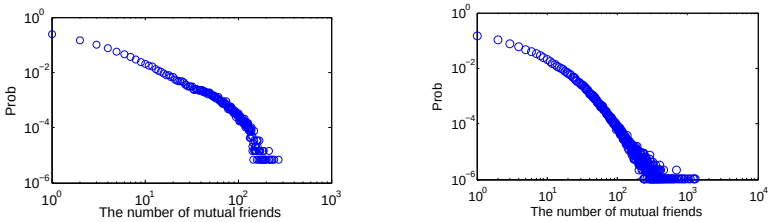
$$PAE_{G,G'} = \frac{|E' - E|}{|E|} \tag{10}$$

$$PRE_{G,G'} = \frac{|E - E'|}{|E|} \tag{11}$$

These metrics are wildly used in the privacy-preserving social network publication to evaluate the performance of privacy-preserving algorithms, and they evaluate the utility of social networks from different aspects. For the first three metrics, we compare the metric value of original social network with the value of anonymized social network, and the closer the value of anonymized network get to the value of original network, the more utility of the original network is maintained, indicating better performances. Under the last metric, percentage of changed edges, smaller values are indicating better performances.



(a) ACM (b) Cora



(c) Brightkite (d) Gowalla

Figure 12: The distribution of NMF

5.3 The Distribution of NMF

The Property 2 in Section 3 states that the distribution of NMF follows a power law or at least asymptotically, i.e., scale free distribution. In Section 4, we use the scale free distribution to design our algorithms. Therefore, we check the distribution of NMF on four dataset in Subsection 5.1. For each pair of friends(an edge between them), we compute their number of mutual friends. Then for each value s of NMFs, we count the number of friend pairs having s mutual friends, indeed get the ratio of friend pairs having s mutual friends. We take the log of ratios and NMFs, and plot the results in Figure 12.

The dataset ACM describes the citation between papers. The mutual friend of two papers means that they have the citing relationship with the same paper. Hence, the number of mutual friends of two papers will be small, and maximum value is 28. The dataset Cora contains a social network representing the collaboration between authors of papers. Some authors may have many common co-authors, i.e., mutual friends, and the maximum number of mutual friends of two authors is 208. The social network Brightkite shows the friendship between users, and the maximum number of mutual friends of two users is 272. The maximum number of mutual friends of two users in the social network Gowalla is 1297. From Figure 12, we can see that the distribution of NMF on four datasets follows the scale free distribution.

5.4 Mutual Friend Attack in Real Data

In the k-degree anonymization model, the adversary re-identifies a vertex using the degree of this vertex. In the k-NMF anonymization model, the adversary re-identifies an edge using the NMF of this edge. We compare both attacks on the real datasets listed in Subsection 5.1, and show the results in Table 1. We removed all labels in four datasets. From Table 1, we can see that the number of edges violating k-NMF anonymity can be sizable when we set k from 5 to 100. It is a very easy way for an adversary to take the mutual friend attack.

Table 1: The numbers of vertices violating k-degree anonymization and edges violating k-NMF anonymization

	ACM		Cora		Brightkite		Gowalla	
k	k-deg	k-NMF	k-deg	k-NMF	k-deg	k-NMF	k-deg	k-NMF
5	54	28	141	106	266	93	467	322
10	75	28	267	179	533	129	721	612
15	103	43	408	277	705	285	1028	858
20	137	62	446	349	795	393	1301	1100
25	162	62	584	488	891	598	1454	1229
30	221	62	752	575	1088	762	1588	1498
50	262	99	1142	733	1425	1297	2504	2108
100	526	226	1472	1350	2326	2578	3799	3511

k-NMF anonymization problem can be seen as a parallel of the k-degree anonymization problem.

5.5 Evaluating k-NMF Anonymization Algorithms

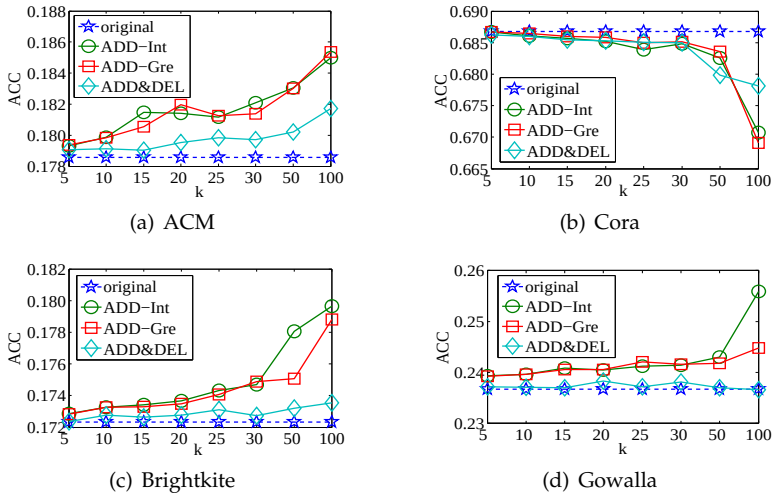
We evaluate the performance of the Greedy and Intuitive ADD algorithms and the ADD&DEL algorithm by measuring the average clustering coefficient, average path length, betweenness centrality and the percentage of changed edges. Figures 13-16 show the results, where ADD-Int and ADD-Gre stand for the ADD algorithm with IntuitGroup and GreedyGroup respectively. ADD&DEL stands for the ADD&DEL algorithm.

Average Clustering Coefficient (ACC): We first compare the average clustering coefficients of the anonymized graphs with the original graph, and the results are shown in Figure 13. The CC values on datasets ACM, Brightkite and Gowalla increase when k increases, but decreased on dataset Cora when k increases. Hence no clear trend on ACC change can be concluded. The analysis on the trend is not our research point. We only focus on the difference of ACC values between original graphs and anonymized graphs. The average clustering coefficients derived by our three methods deviate slightly from the original values on four datasets. The ADD&DEL performs better than the two ADD algorithms in Figure 13, and the ADD algorithm with GreedyGroup looks slightly better than the algorithm with IntuitGroup.

Average Path Length (APL): Figure 14 shows the average path lengths for the anonymized graphs and the original graphs on four datasets. The APL of the graph anonymized by the ADD&DEL algorithm is very close to the APL of the original graph. By adding and deleting edges, the ADD&DEL algorithm can preserve more utility than the ADD algorithm. Besides, the differences of APL between the graphs anonymized by our methods and the original graphs are very small, and the largest difference value is 0.8 when k is set as 100 on the dataset Cora.

Betweenness Centrality (BC): All the plots of the average betweenness centralities are very similar to the plots of the *APL*. Hence we show the distribution of betweenness centralities of all vertices in Figure 15. The sub-figures in Figures 15(a), 15(b) and 15(c) enlarge the details on the frequency varied from 0 to 100. Clearly, in Figures 15(a) and 15(b), ADD&DEL performs better than the ADD algorithm with GreedyGroup, and shows little sensitivity to the value of k while ADD with GreedyGroup degrades as k increases. Also Figure 15(c) shows that ADD&DEL performs better than the ADD algorithms. Figure 15 shows that the difference of *BC* between the anonymized graph and the original graph is very small.

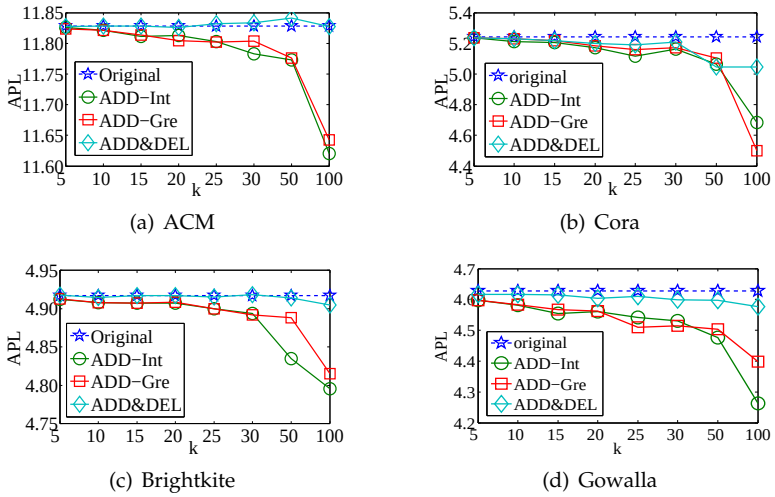
Percentage of changed edges (PCE): As there is no vertex addition occurred in all cases



(a) ACM (b) Cora

(c) Brightkite (d) Gowalla

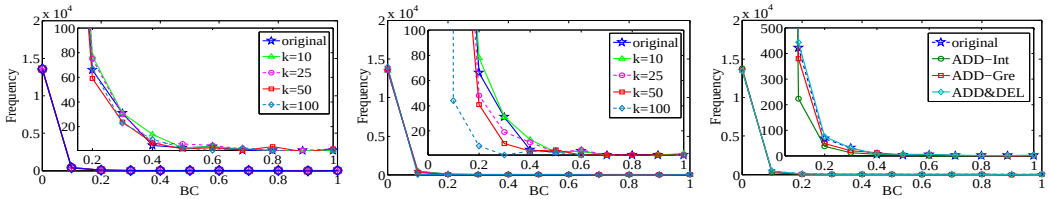
Figure 13: Clustering coefficients



(a) ACM (b) Cora

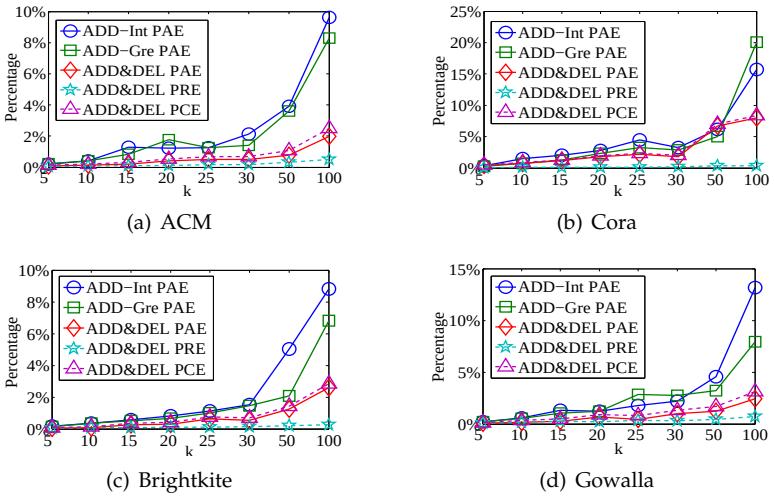
(c) Brightkite (d) Gowalla

Figure 14: Average path lengths



(a) ADD&DEL algorithm (b) ADD algorithm with Greedy-Group (c) k is 25

Figure 15: Betweenness centrality distributions on Cora



(a) ACM (b) Cora

(c) Brightkite (d) Gowalla

Figure 16: Edge changes

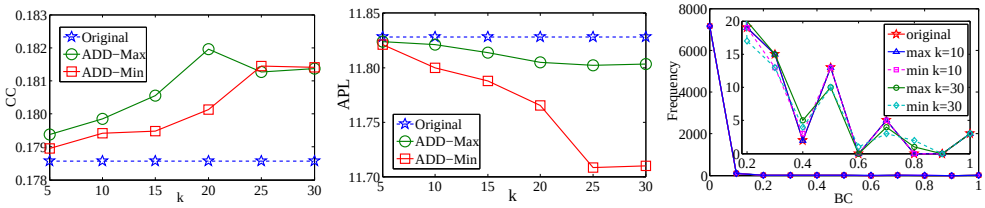
considered under ADD and ADD&DEL which do not perform node deletion operations, we consider the edge changes. Figure 16 shows the edge changes on the original graphs. The changes on ADD&DEL includes the percentages of added edges (PAE) and removed edges (PRE). The ADD&DEL algorithm changed fewest edges, and the ADD algorithm with GreedyGroup added fewer edges than the algorithm with IntuitGroup.

From the above evaluation, we can see that our algorithms can preserve the utility of the original graph effectively. Among them, ADD&DEL performs better than the ADD algorithm, and GreedyGroup performs better than IntuitGroup.

5.6 Verifying the Candidate Vertex Selection

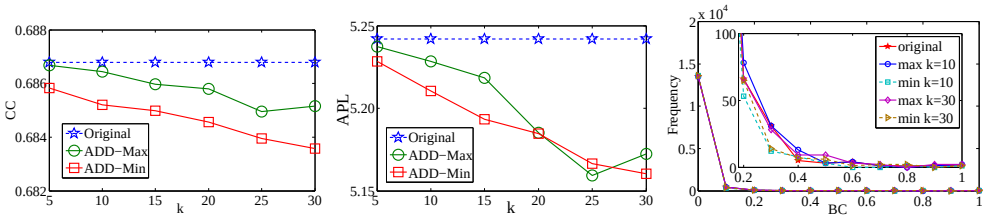
In the ADD algorithm, a vertex v tends to link the candidate vertex w_{max} which has the most mutual friends with v among all the candidate vertices. In order to validate why we select w_{max}, we compare the results of the ADD algorithm with the ADD-min algorithm, which tends to link the vertex w_{min} having the fewest mutual friends with v. Figures 17 and 18 show the results on the datasets ACM and Cora. Similar trend is observed on Brightkite and Gowalla.

In Figures 17(a) and 18(a), we show the CC values of the graphs anonymized by the ADD algorithm and the ADD-min algorithm. The ADD-max and ADD-min curves represent the results of the ADD algorithm and the ADD-min algorithm respectively. The ADD-max curve is closer to the original curve than the ADD-min curve on dataset Cora. As the trend of CC value changes is not clear, the ADD-min curve may sometime be more closer than ADD-max. The ADD-max curve is closer to the original curve than the ADD-min curve on the datasets ACM and Cora in Figures 17(b) and 18(b). Finally we compare the results on the distribution of betweenness centrality in Figures 17(c) and 18(c). The solid lines and dashed lines represent the results of the ADD algorithm and the ADD-min algorithm respectively. The solid lines are closer to the lines of the original graphs than dashed lines in the both figures. These results reflect that choosing the vertex w_{max} is reasonable for anonymizing the original graphs.



(a) CC (b) APL (c) BC

Figure 17: Utility preservation of dataset ACM



(a) CC (b) APL (c) BC

Figure 18: Utility preservation of dataset Cora

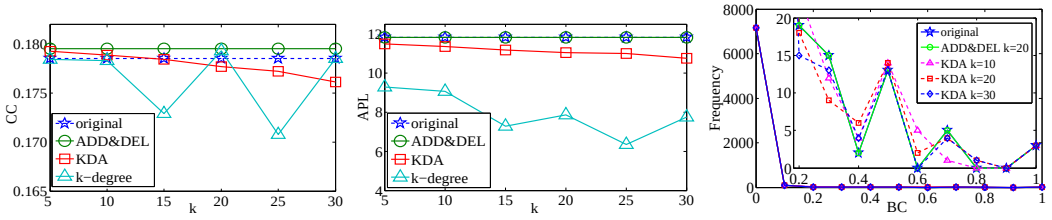
5.7 Evaluating the *KDA* Algorithm

In this subsection, we evaluate the performance of the KDA algorithm in Section 4.4 , and compare it with the classic k-degree anonymization algorithm in [22].

Since there are no new triangles formed after the KDA algorithm adds new edges, the clustering coefficient decreases a little bit as k increases as shown in Figures 19(a), 20(a) and 21(a). Our algorithm performs better than the classic k-degree anonymization on this measure. Since new edges are added into the graph, the APL value decreases a little bit as k increases as shown in Figures 19(b), 20(b), and 21(b). As we consider the k-NMF anonymity, the classic k-degree anonymization performs a little better than our algorithm on the APL measure. But when the APL of the graph is large, our algorithm can perform better than the classic k-degree anonymization as shown in Figure 19(b). The results show that our algorithm performs well on preserving the utility while protecting the privacy by carefully exploring the graph property. The classic k-degree anonymization makes less effort on this except minimizing the number of edges added. Figures 19(c), 20(c) and 21(c) show the distributions of betweenness centrality of graphs anonymized by the KDA algorithm when we set k_{deg} as 10, 20 and 30. The distributions of the anonymized graphs are very similar to the distributions of the original graphs especially for the ACM and Brightkite datasets. It shows that the KDA algorithm can preserve much of the utility of the graph anonymized by the k-NMF algorithms.

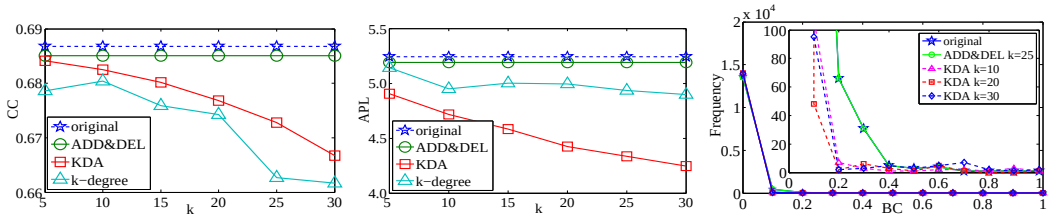
6 Conclusions

In this paper, we have identified a new problem of k-anonymity on the number of mutual friends, which protects against the mutual friend attack in the social network publication. To solve this problem, we designed two heuristic algorithms which consider the utility of



(a) CC (b) APL (c) BC

Figure 19: k-degree anonymization on 20-NMF anonymized graph of ACM



(a) CC (b) APL (c) BC

Figure 20: k-degree anonymization on 25-NMF anonymized graph of Cora

the graph. We also devised an algorithm to ensure the k-degree anonymity based on the k-NMF anonymity. The experimental results demonstrate that our approaches can ensure the k-NMF anonymity while preserve much of the utility in the original social networks.

References

[1] Stanford Network Analysis Project. Available online: http://snap.stanford.edu/index.html. 2007.

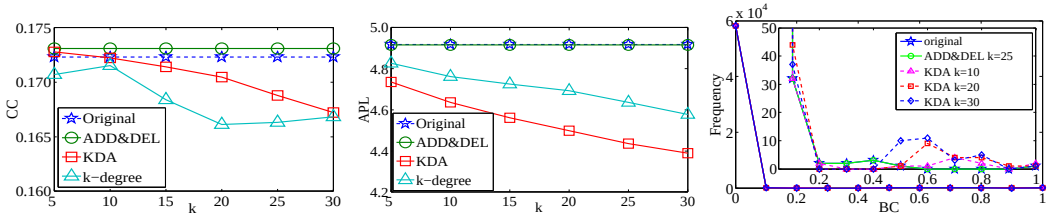
[2] B. Zhou, J. Pei, and W. Luk, "A brief survey on anonymization techniques for privacy preserving publishing of social network data," *ACM SIGKDD Explorations Newsletter*, vol. 10, no. 2, pp. 12–22, 2008.

[3] X. Wu, X. Ying, K. Liu, and L. Chen, "A survey of privacy preservation of graphs and social networks," *Managing and Mining Graph Data*, vol. 40, pp. 421–453, 2010.

[4] C.-H. Tai, P. S. Yu, D.-N. Yang, and M.-S. Chen, "Privacy preserving social network publication against friendship attacks," in *Proc. of KDD, San Diego, CA*, 2011, pp. 1262–1270.

[5] P. Samarati, L. Sweeney. "Protecting Privacy When Disclosing Information: k-Anonymity and Its Enforcement through Generalization and Suppression[R]," Palo Alto, CA: SRI Computer Science Laboratory, 1998.

[6] A. Machanavajjhala, D. Kifer, J. Gehrke, et al. "l-Diversity: Privacy Beyond k-Anonymity[J]," *ACM Transactions on Knowledge Discovery from Data*, Vol. 1, pp. 1–47, 2007.



(a) CC (b) APL (c) BC

Figure 21: k-degree anonymization on 25-NMF anonymized graph of Brightkite

[7] N. Li, T. Li, S. Venkatasubramanian. "t-Closeness: Privacy beyond k-Anonymity and l-Diversity," in *Proc. of the 23rd International Conference on Data Engineering, Istanbul, Turkey*, 2007, pp. 106–115.

[8] M. E. Nergiz, M. Atzori, C. Clifton. "Hiding the Presence of Individuals from Shared Databases," in *Proc. of the ACM SIGMOD International Conference on Management of Data, Beijing*, 2007, pp. 665–676.

[9] B. Fung, K. Wang, R. Chen, et al. "Privacy-Preserving Data Publishing: A Survey of Recent Developments," *ACM Computing Surveys (CSUR)*, 2010, Vol. 42, no. 4, Article No. 14.

[10] M. Hay, G. Miklau, D. Jensen, D. Towsley, and P. Weis, "Resisting Structural Re-Identification in Anonymized Social Networks, *Proc. VLDB Endowment*, vol. 1, pp. 102-114, 2008.

[11] A. Campan and T. M. Truta, "A Clustering Approach for Data and Structural Anonymity in Social Networks, in *Proc. Second ACM SIGKDD Intl Workshop Privacy, Security, and Trust in KDD (PinKDD 08)*, 2008.

[12] E. Zheleva and L. Getoor, "Preserving the Privacy of Sensitive Relationships in Graph Data, in *Proc. First SIGKDD Intl Workshop Privacy, Security, and Trust in KDD (PinKDD 07)*, pp. 153–171, 2007.

[13] G. Cormode, D. Srivastava, T. Yu, and Q. Zhang, "Anonymizing Bipartite Graph Data Using Safe Groupings, in *Proc. VLDB Endowment*, vol. 1, pp. 833-844, 2008.

[14] S. Bhagat, G. Cormode, B. Krishnamurthy, and D. Srivastava, "Class-Based Graph Anonymization for Social Network Data,in *Proc. VLDB Endowment*, vol. 2, pp. 766-777, 2009.

[15] A. Campan, T. M. Truta, and N. Cooper, "P-Sensitive K-Anonymity with Generalization Constraints, *Trans. Data Privacy*, vol. 2, pp. 65-89, 2010.

[16] M. Hay, G. Miklau, D. Jensen, et al. "Anonymizing Social Networks," Amherst: University of Massachusetts, 2007.

[17] X. Ying, X. Wu. "Randomizing Social Networks: A Spectrum Preserving Approach," in *Proc. of the 8th SIAM International Conference on Data Mining, Atlanta, GA*, 2008, pp. 739–750.

[18] S. Hanhijärvi, G. C. Garriga, K. Puolamäki. "Randomization Techniques for Graphs," in *Proc. of the 9th SIAM International Conference on Data Mining, Sparks, NV*, 2009, pp. 780–791.

[19] X. Ying, X. Wu. "Graph Generation with Prescribed Feature Constraints," in *Proc. of the 9th SIAM International Conference on Data Mining, Sparks, NV*, 2009, pp. 966–977.

[20] L. Liu, J. Wang, J. Liu, et al. "Privacy preserving in social networks against sensitive edge disclosure,". Technical Report Technical Report CMIDA-HiPSCCS 006-08, Department of Computer Science, University of Kentucky, KY, 2008.

[21] L. Backstrom, C. Dwork, and J. Kleinberg, "Wherefore art thou r3579x? anonymized social networks, hidden patterns and structural steganography," in *Proc. of WWW, Banff, Alberta*, 2007, pp. 181–190.

[22] K. Liu and E. Terzi, "Towards identity anonymization on graphs," in *Proc. of SIGMOD, Vancouver, BC*, 2008, pp. 93–106.

[23] B. Zhou and J. Pei, "Preserving privacy in social networks against neighborhood attacks," in *Proc. of ICDE, Cancun, Mexico*, 2008, pp. 506–515.

[24] J. Cheng, A. W. Fu, and J. Liu, "K-isomorphism: privacy preserving network publication against structural attacks," in *Proc. of SIGMOD, Indianapolis, IN*, 2010, pp. 459–470.

[25] L. Zou, L. Chen, and M. T. Özsu, "K-automorphism: a general framework for privacy preserving network publication," *Journal Proceedings of the VLDB Endowment*, vol. 2, no. 1, pp. 946–957, 2009.

[26] W. Wu, Y. Xiao, W. Wang, Z. He, and Z. Wang, "K-symmetry model for identity anonymization in social networks," in *Proc. of EDBT, Lausanne, Switzerland*, 2010, pp. 111–122.

[27] M. Hay, G. Miklau, D. Jensen, D. Towsley, and P. Weis, "Resisting structural re-identification in anonymized social networks," *The VLDB Journal*, vol. 19, no. 6, pp. 797–823, 2008.

[28] B. Zhou, J. Pei. "The k-Anonymity and l-Diversity Approaches for Privacy Preservation in Social Networks against Neighborhood Attacks," *Knowledge and Information Systems*, 2011, vol. 28, no. 1, pp. 47–77.

[29] C.-H. Tai, P. S. Yu, D.-N. Yang, and M.-S. Chen, "Structural diveristy for pricacy in publishing soical networks," in *Proc. of SDM, Mesa, AZ*, 2011, pp. 35–46.

[30] L. Zhang, and W. Zhang. "Edge anonymity in social network graphs," In *Int'l Conf. on Computational Science and Engineering*, 2009, pp. 1–8.

[31] E. Zheleva, and L. Getoor. "Preserving the privacy of sensitive relationships in graph data," In *Privacy, security, and trust in KDD*, pp. 153–171. 2008.

[32] M. Hay, C. Li, G. Miklau, and D. Jensen, "Accurate estimation of the degree distribution of private networks," in *ICDM*, 2009, pp. 169–178.

[33] V. Rastogi, M. Hay, G. Miklau, and D. Suciu, "Relationship privacy: Output perturbation for queries with joins," in *PODS*, 2009, pp. 107–116.

[34] Y. Wang, X. Wu, J. Zhu, and Y. Xiang, "On learning cluster coefficient of private networks," in *ASONAM*, 2012.

[35] Y. Wang, X. Wu, and L. Wu, "Differential privacy preserving spectral graph analysis," in *PAKDD*, 2013.

[36] D. Mir and R.Wright, "A differentially private graph estimator," in *ICDMW*, 2009, pp. 122–129.

[37] A. Sala, X. Zhao, C.Wilson, H. Zheng, and B. Zhao, "Sharing graphs using differentially private graph models," in *SIGCOM*, 2011, pp. 81–98.

[38] P. Mahadevan, D. Krioukov, K. Fall, and A. Vahdat, "Systematic topology analysis and generation using degree correlations," in *ACM SIGCOMM Computer Communication Review*, vol. 36, no. 4, 2006, pp. 135–146.

[39] Y. Wang and X. Wu. "Preserving Differential Privacy in Degree-correlation based Graph Generation". *Transactions on Data Privacy*, Vol. 6, No. 2, 2013, pp. 127–145.

[40] M. Faloutsos, P. Faloutsos, and C. Faloutsos, "On power law relationships of the internet topology," in *Proc. of SIGCOMM, Cambridge, MA*, 1999, pp. 251–262.

[41] V. Zlatic, D. Garlaschelli, and G. Caldarelli, "Complex networks with arbitrary edge multiplicities," *Physics*, vol. 97, pp. 8–11, 2011.

[42] D. Liben-Nowell and J. Kleinberg, "The link-prediction problem for social networks," *Journal of the American society for information science and technology*, vol. 58, no. 7, pp. 1019–1031, 2007.

[43] A. K. McCallum, K. Nigam, J. Rennie, and K. Seymore, "Automating the construction of internet portals with machine learning," *Information Retrieval Journal*, vol. 3, pp. 127–163, 2000.