A Local-Perturbation Anonymizing Approach to Preserving Community Structure in Released Social Networks

Huanjie Wang, Peng Liu, Shan Lin, and Xianxian Li(✉)

Guangxi Key Lab of Multi-source Information Mining and Security,
Guangxi Normal University, Guilin 541004, China
whj.6040@163.com, lin-sam@foxmail.com, {liupeng,lixx}@gxnu.edu.cn

Abstract. Social networks provide a large amount of social network data, which is gathered and released for various purposes. Since social network data usually contains much sensitive information of individuals, the data needs to be anonymized before releasing. To protect privacy of individuals in released social network, many anonymizing methods have been proposed. However, most of them were proposed for general purpose, and suffered the over-information loss problem when they were used for specific purposes. In this paper, we focus on the problem of preserving structure information in anonymized social network data, which is the most important knowledge for community analysis. Furthermore, we propose a novel local-perturbation technique that can reach the same privacy requirement of k-anonymity, while minimizing the impact on community structure. We evaluate the performance of our method on real-world data. Experimental results show that our method has less community structure information loss compared with existing techniques.

Keywords: Social networks · Privacy protection · Community structure

1 Introduction

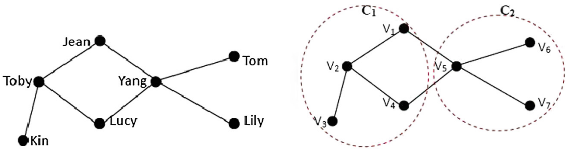
Recently, social network applications have provided a large amount of information, which is increasing continually and has more value for data analysis, such as researching the cause of social phenomenon [7], etc. However, we could not release social network data in raw form, which can raise serious privacy concerns, because it contains sensitive information. In this paper, we present a method to anonymize social network data for preventing individuals from re-identifying, while achieving the maximum utility of community structure for analysis.

1.1 Motivation

To protect privacy of individuals, a naive anonymizing method is proposed by removing the unique identifies of nodes. However, it is insufficient and has been discussed in previous work [9].

J.-H. Lee and S. Pack (Eds.): QShine 2016, LNICST 199, pp. 36–45, 2017.
DOI: 10.1007/978-3-319-60717-7_4

Example 1. A typical social network is presented in Fig. 1a, and its naive anonymized graph is depicted in Fig. 1b by removing names of participants. Even so, an adversary could re-identify the target victim by some more complex structure attack [4]. Assume that the adversary knows that Kin has one friend, and his friend has three friends. It is easy to infer that Kin is V_3 in Fig. 1b.



(a) (b)

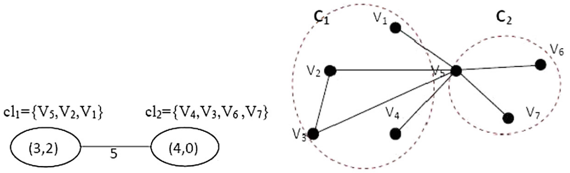
Fig. 1. (a) a raw social network (b) a naive anonymized social network

An effective way to protect individual privacy in Example 1 is k-anonymity [2,3], which divides all nodes into several clusters, and each cluster has at least k indistinguishable nodes. These clusters are generalized to super nodes, and the edges among clusters are generalized to super edges. However, because most of graph analysis methods can only process atomic nodes and edges, the k-anonymity graph usually is reconstructed before analyzing [2,3].

By reason of do not considering the community structure information in the clustering process [2,3], the boundaries of original community structure are likely to become blurry after reconstructing.

Example 2. As shown in Fig. 1b, there are 2 communities $\{C_1, C_2\}$ and 2 edges between them. Figure 2a shows its 3-anonymity graph, and Fig. 2b is a possible result of reconstruction. Then, the number of edges connecting C_1 and C_2 becomes 4, which blurs the boundary between them seriously. Obviously, for the community structure information, there is a big difference to the original graph.

To address the problems above, we propose a novel local-perturbation approach, which can achieve the same requirement of the k-anonymity, while preserving "high" utility of the community structure for data releasing, so that data analyzers could take some relative researches about community structure.



(a) (b)

Fig. 2. (a) a 3-anonymity graph (b) a reconstructed graph

1.2 Contributions

The contributions of this paper are summarized as follows.

1. We propose an anonymizing approach that is designed for preserving the community structure in released social network data. This is a crucial difference to existing methods in [2], which just analyzed how well the communities in social networks were preserved by using existing anonymous techniques.
2. By combining the clustering technique with the randomly reconstructing technique, we propose a novel local-perturbation approach to reaching the same privacy level of k-anonymity, while minimizing the impact on community structure. Experiment results demonstrate that our approach can effectively preserve the privacy with the reasonable trade-off between privacy and data utility measured in terms of preserving community structure.

The rest of the paper is organized as follows. Section 2 presents related work of anonymization. The problem is defined in Sect. 3. Our anonymizing method is described in Sect. 4. The concrete evaluation criterions and the experimental results are discussed in Sect. 5, and Sect. 6 concludes this paper.

2 Related Work

There are considerable research efforts for designing privacy-preserving methods in social network. The privacy of social network data can be mainly categorized into two types. One type is node-privacy, in which many researches mostly focus on node re-identification [4] and nodes' attribute disclosure [3]. For node re-identification, the attack goal is to identify the target victim for achieving more beneficial information; and for nodes' attribute disclosure, the attack goal is to infer sensitive information of target victim, such as disease and salary. The other type is edge-privacy, which contains link re-identification [11] and edges' attribute disclosure [3]. For link re-identification, the attack goal is to identify sensitive relationships between nodes; and for edges' attribute disclosure, the attack goal is to infer some sensitive relationship categories between nodes. This paper focuses on preventing the node from re-identifying in unlabeled graph.

In order to protect the sensitive information mentioned above, some anonymous techniques have been proposed in these years. These techniques can be classified into four categories: adding nodes [5,8], adding and deleting edges [4], generalization [3], and randomization [1]. In this paper, we combine clustering technique with randomly reconstructing technique, which can get a local-perturbation anonymized graph that has the same number of edges and nodes as the original graph.

Recently, the researches about the node re-identification in the community have been studied [12]. Tai et al. [12] presented the model of structural diversity, for each node v, there must exist at least $k-1$ other nodes located in at least $k-1$ other communities with the identical degree of v.

On the whole, some studies related to this paper are social network clustering model and reconstructing model. Besides, we also use the community detection approach to detecting the community structure of social network graph.

3 Problem Descriptions

In this paper, we model an initial social network as an undirected graph $G = (V, E)$, where V is a set of nodes without labels, and $E \in V \times V$ is a set of edges without labels. Each node indicates to an individual in the underlying group. An edge between two nodes presents the relationship between the two corresponding individuals. Only binary relationships are allowed in our model.

3.1 The Privacy Model

Suppose that an adversary knows any subgraph information of the target victim location, and wants to re-identify the target victim node in the released data. The problem in this paper is how to transform a given social network G into an anonymous graph G', which satisfies the requirement of k-anonymity [2,3], while preserving community structure information as much as possible.

Definition 1 (k-anonymity social network). Let G be a social network and G' be an anonymization of G. If G' is k-anonymity, then with any subgraph background knowledge, any node in G cannot be re-identified in G' with confidence larger than $1/k$.

3.2 Relevant Definitions

The nodes in the social network always tend to form closely-knit groups, these groups are also known as communities.

Definition 2 (communities in social network). Let $G = (V, E)$ is a social network, the set of communities $C = \{C_1, C_2, \cdots, C_m\}$, where $C_i \cap C_j = \phi$ for all $1 \leq i \neq j \leq m$. For each $C_o \in C$, the density of internal connection is higher than outside.

In order to protect the community structure, we choose a classic community detection GN algorithm [10] to discover community structure of the original network, which uses the modularity [6,10] optimization method that is defined as

$$Q = \sum_{c=1}^{n} [\frac{l_c}{m} - (\frac{d_c}{2m})^2] \tag{1}$$

where n is the number of communities, l_c is the total number of edges in the community c, d_c is the sum of degrees of nodes in c, m is the number of edges in G.

Our technique mainly includes two processes, clustering and reconstruction. Then, some relevant concepts about anonymization are defined as follows.

Definition 3 (k-cluster social network). Let $G = (V, E)$ is a social network, and k is a threshold specified by social network data holder. For a given clustering

$CL = \{cl_1, cl_2, \cdots, cl_n\}$ of V, the corresponding social network is denoted as G_{cl} where $cl_t \cap cl_c = \phi$ for all $1 \leq t \neq c \leq n$, and $|cl_i| \geq k$ for $1 \leq i \leq n$.

In the clustering process, the shortest distance is important evidence. We use the symbol $adj[V_i]$ to denote the set of neighbors of a node V_i. The distance between nodes is defined as follow.

Definition 4 (the distance between nodes). The distance between two nodes (V_i, V_j) is

$$dist(V_i, V_j) = \frac{|\{V_k | V_k \in (adj[V_i] \bigoplus adj[V_j]), V_k \neq V_i \neq V_j\}|}{n-2} \qquad (2)$$

where n is the number of nodes in graph. The reason that n is reduced by 2 in the denominator is that we exclude V_i and V_j from the set. For example, the distance between V_1 and V_2 in Fig. 1b is 3/5.

Then, we will get the distance between a node and a cluster [2,3].

Definition 5 (the distance between a node and a cluster). The distance between a node V_p and a cluster cl_q is

$$dist(V_p, cl_q) = \frac{\sum_{V_j \in cl_q} dist(V_p, V_j)}{|cl_q|} \qquad (3)$$

3.3 Problem Statement

In this paper, we address the following problem.

Definition 6 (social network anonymization for community structure). Given a social network G without labels, and a privacy requirement k. The problem of social network anonymization for community structure is to transform G to a local-perturbation social network *G'*, which satisfies the given anonymous requirement while preserving community structure as much as possible.

4 The Anonymization Method

In this section, we introduce a method to transform the original social network G into a local-perturbation graph *G'* for privacy preservation, and achieve the maximum of the utility of community structure. The first step is to transform G into a k-cluster graph G_{cl}, and then reconstruct each cluster in G_{cl}.

4.1 Cluster for Social Networks

Owing to the fact that optimal clustering problem is known to be NP-hard [3], we devise a greedy clustering approach named K-Cluster presented in Table 1 that is based on *SaNGreeA* (Social Network Greedy Anonymization) algorithm [2,3], and the time complexity is also same as [2,3].

Table 1. Algorithm 1 K-Cluster(G)

```
Input: Social network $G = (V, E)$ and the threshold $k$
Output: A $k$-cluster graph $G_{cl}$ and number $n$ of clusters
1: $CL = \phi$; $i = 1$; $m = 1$;//sort $V$ by degree in descending order
2: while $|V| >= k$ do
3:     $V^i_{seed} = V[0]$; $cl_i = \{V^i_{seed}\}$; $V = V - cl_i$; //select the seed node for $cl_i$
4:     while $|cl_i| < k$: $FindBestNode(V, cl_i)$;
5:     $CL = CL \cup \{cl_i\}$; $i + +$;
6: end while
7: if $V \neq \phi$ :    // find the best cluster for each of them
8:     for every $V_m$ in in current V:
9:         $FineBestCluster(V_m, CL)$;
10: end if
```

Before anonymizing, all nodes in V should be sorted by degree in descending order. A new cluster is formed with a node in current V that has the maximum degree (Line 3). Then the algorithm gathers nodes one by one to this current cluster until it has k nodes (Line 4). Due to the power law degree distribution, it is likely that more than one node have the same degree, and this may result in that there are more than one node have the same distance from current cluster. The question is how to select the proper nodes from these candidates for the current cluster with minimal impact on community structure. Different selections lead to different results. Thus, we devise a heuristic algorithm presented in Table 2 for selecting proper nodes.

Table 2. Function 1 $FindBestNode(V, cl_i)$

```
for each node $V_p$ in $V$: compute $dist(V_p, cl_i)$;
  get all $V_p$ from the minimum distance and named them as candidate set $CanN$;
if $|CanN| > 1$ :
    for each node $V_q$ in $CanN$://traverse the list
        if $V_q$ and $V^i_{seed}$ are in the same community:
            add the $V_q$ to $cl_i$, and $V = V - \{V_q\}$;break;
    if there is no such node, add the last node $V_q$ to $cl_i$, and $V = V - \{V_q\}$;
else: add the $V_q$ in $CanN$ to $cl_i$, and $V = V - \{V_q\}$;
```

Besides, when the number of nodes in G is not a multiple of k, it is possible that the number of nodes in current V is less than k. Then, we should find the best cluster for each of them. The specific technique is described as Table 3.

4.2 Reconstruction

To protect user's privacy and analyze data conveniently, the k-cluster social network must to be reconstructed before releasing. However, it will bring more

Table 3. Function 2 $FindBestCluster(V_m, CL)$

```
$mincl = dist(V_m, cl_0)$;
for each cluster $cl_n$ in $CL$: // $1 \leq n \leq i-1$
    if $V_m$ and $V_{seed}^i$ are in the same community:
        $Bestcl = cl_n$; Break;
    else: compute $dist = dist(V_m, cl_n)$;
        if $dist < mincl$: $mincl = dist$; $Bestcl = cl_n$;
add $V_m$ to $Bestcl$, $cl_i = cl_i - \{V_m\}$;
```

uncertainty to reconstruct the entire graph, which is worse for data analyzers to achieve accurate community structure information. Here, we will reconstruct k-cluster graph by randomly regenerating edges in each cluster uniformly and make sure that the number of intra-cluster edges in each cluster is the same as before. Besides, the inter-cluster edges stay the same as before.

By the reason of uniform probability of selecting any pair nodes to regenerate edges in each cluster during the reconstructing process, the probability of each node is selected is equal, in other words, the nodes in one cluster are indistinguishable. Besides, the size of each cluster is no less than k, therefore, the probability for an adversary re-identifies any node in the anonymized social network *G'* is no more than $1/k$. Then we can safely get that our local-perturbation approach could achieve the same privacy requirement of k-anonymity.

5 Experimental Evaluations

To evaluate the effectiveness of our algorithm, we compare our local-perturbation algorithm with the *SaNGreeA-uniform* anonymizing algorithm proposed in [2,3].

5.1 Datasets and Data Utility

We study the data utility on three real datasets [8]:

- *WebKB* (http://linqs.umiacs.umd.edu/projects//projects/lbc/index.html).
- *Citation* (http://www.datatang.com/data/17310).
- *Cora* (http://www.cs.umd.edu/projects/linqs/projects/lbc/index.html).

We use Jaccard similarity and the change of modularity $\triangle Q$ to compare the results of community structure preservation between initial graph and anonymized graph.

Firstly, we consider the Jaccard similarity. The GN algorithm can detect communities for the initial network and the anonymized network. The set of communities before amonymizing is represented as $C = \{C_1, C_2, \cdots, C_n\}$, and after anonymizing is represented as $C = \{C'_1, C'_2, \cdots, C'_m\}$.

$$J_i(C_i) = \frac{|C_i \cap C'_j|}{|C_j \cup C'_j|}, i \in [1, n], j \in [1, m] \quad (4)$$

The integral community structure preservation based on Jaccard similarity is computed as the average of all $J_i(C_i)$.

$$J(G, G') = \frac{\sum_{i=1}^{n} J_i}{n}, i \in [1, n] \tag{5}$$

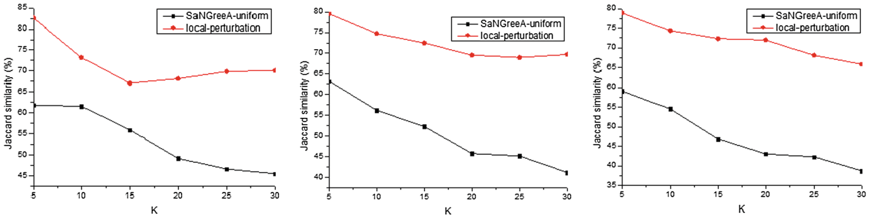
In addition, we also use the change of modularity to compare the community information preservation level. Intuitively, the greater the result gets, the more of the boundaries between communities become blurry, that is, the community structure information of original data does not get better preservation.

$$\triangle Q = Q - Q' \tag{6}$$

5.2 Results and Analysis

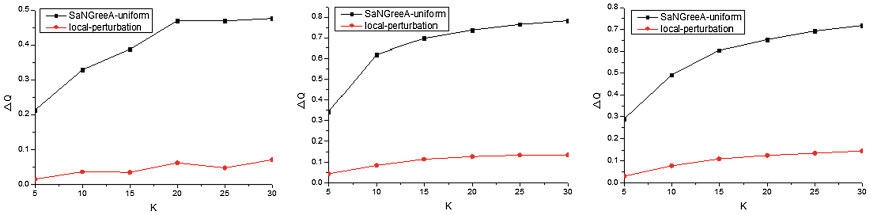
Firstly, we evaluate the impact of anonymization on community structure, and the data utility is calculated by the Jaccard similarity and $\triangle Q$.

Figures 3 and 4 represent Jaccard similarity and $\triangle Q$ in terms of changing k values using *SaNGreeA-uniform* algorithm and local-perturbation algorithm respectively. The former figure shows the community structure of original social network has more serious damage with the increase of the values of k. However, our method has more obvious advantages on community structure protection.



(a) WebKB (b) Citation (c) Cora

Fig. 3. Jaccard similarity for different k



(a) WebKB (b) Citation (c) Cora

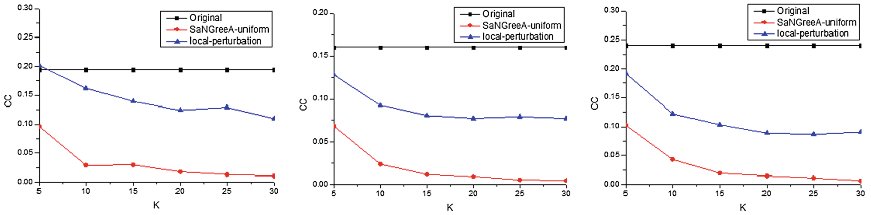
Fig. 4. $\triangle Q$ for different k

The latter figure suggests that the boundaries between communities of original social network become more blurry with the increase of k, and our algorithm has a relatively smaller impact than *SaNGreeA-uniform* algorithm, because we preserve more community structure information by using our technique.

5.3 Other Structural Property Analysis

The social network is a complex data structure and has many topological properties. In addition to contrast the impact of anonymization for community structure, the average clustering coefficient (CC) is also evaluated.

The change of CC is presented in Fig. 5. With the increase of k, CC becomes smaller and smaller after anonymizing and CC values of *SaNGreeA-uniform* algorithm are even close to 0. Intuitively, our approach has lower differences to the original data.



(a) WebKB (b) Citation (c) Cora

Fig. 5. CC for different k

6 Conclusion

In this paper, we formally define the problem of social network anonymization for releasing, and propose a novel local-perturbation approach that combines clustering technique with randomly reconstructing technique to transform the original network to the released network. Because of considering the community structure in anonymous procedure, our proposed technique can reach the same privacy requirement of k-anonymity, while minimizing the impact on community structure. We performe experiments on three real datasets with three measurements and demonstrate that our method can provide the same privacy protection level of k-anonymity and have less community structure information loss compared with existing techniques.

Acknowledgments. The research is supported by the National Science Foundation of China (Nos. 61272535, 61502111), Guangxi "Bagui Scholar" Teams for Innovation and Research Project, the Guangxi Collaborative Center of Multi-source Information Integration and Intelligent Processing, Guangxi Natural Science Foundation (Nos. 2013GXNSFBA019263, 2014GXNSFAA118018 and 2015GXNSFBA139246), and the Science and Technology Research Projects of Guangxi Higher Education (Nos. 2015YB032 and 2013YB029), and the Innovation Project of Guangxi Graduate Education (No. YCSZ2015104).

References

1. Boldi, P., et al.: Injecting uncertainty in graphs for identity obfuscation. Proc. VLDB Endow. **5**(11), 1376–1387 (2012)
2. Campan, A., et al.: Preserving communities in anonymized social networks. Trans. Data Priv. **8**(1), 55–87 (2015)
3. Campan, A., Truta, T.M.: Data and structural *k*-anonymity in social networks. In: Bonchi, F., Ferrari, E., Jiang, W., Malin, B. (eds.) PInKDD 2008. LNCS, vol. 5456, pp. 33–54. Springer, Heidelberg (2009). doi:10.1007/978-3-642-01718-6_4
4. Cheng, J., et al.: K-isomorphism: privacy preserving network publication against structural attacks. In: Proceedings of the 2010 ACM SIGMOD International Conference on Management of data. ACM (2010)
5. Chester, S., et al.: k-anonymization of social networks by vertex addition. In: ADBIS (2) (2011)
6. Fortunato, S.: Community detection in graphs. Phys. Rep. **486**(3), 75–174 (2010)
7. Hechter, M.: Principles of Group Solidarity. University of California Press, Berkeley (1988)
8. Jiao, J., Liu, P., Li, X.: A personalized privacy preserving method for publishing social network data. In: Gopal, T.V., Agrawal, M., Li, A., Cooper, S.B. (eds.) TAMC 2014. LNCS, vol. 8402, pp. 141–157. Springer, Cham (2014). doi:10.1007/978-3-319-06089-7_10
9. Narayanan, A., Shmatikov, V.: De-anonymizing social networks. In: 2009 30th IEEE Symposium on Security and Privacy. IEEE (2009)
10. Newman, M.E., Girvan, M.: Finding and evaluating community structure in networks. Phys. Rev. E **69**(2), 026113 (2004)
11. Tai, C.-H., et al.: Privacy-preserving social network publication against friendship attacks. In: Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM (2011)
12. Tai, C.-H., et al.: Structural diversity for resisting community identification in published social networks. IEEE Trans. Knowl. Data Eng. **26**(1), 235–252 (2014)