Journal of Network and Computer Applications 195 (2021) 103214

Contents lists available at ScienceDirect

Journal of Network and Computer Applications

journal homepage: www.elsevier.com/locate/jnca

Graph partition based privacy-preserving scheme in social networks

Hongyan Zhang [a,b], Limei Lin [a], Li Xu [a,*], Xiaoding Wang [a]

[a] College of Mathematics and Informatics, Key Laboratory of Network Security and Cryptology, Fujian Normal University, Fuzhou, Fujian, 350117, PR China
[b] Concord University College Fujian Normal University, Fuzhou, Fujian, 350117, PR China

ARTICLE INFO

Keywords:
Social Networks
K-anonymity
1-Neighborhood attack
Graph partition
Privacy preservation

ABSTRACT

With the development of social networks, more and more data about users are released on social platforms such as Facebook, Enron, WeChat, in terms of social graphs. Without the efficient anonymization, the graph data publishing will cause serious privacy leakage of users, for example, malicious attackers might launch 1-neighborhood graph attack on targets, which assumes that 1-hop neighbors and the relations among them are known by attackers, thereby, targets can be re-identified in anonymous social graphs. To prevent such attack, we propose a Graph Partition based Privacy-preserving Scheme, named GPPS,i n social networks to realize social graph anonymization. The proposed GPPS preserves users' identity privacy by *k*-anonymity which achieved by node clustering and graph modification. Specifically, in the similarity matrix calculation, we introduce the degree-based graph entropy to improve the accuracy of node clustering. Then, the graph modification is implemented to achieve the *k*-anonymity of users and meanwhile minimize the graph information loss. The experiment results illustrate that the proposed GPPS is effective and efficient both on synthetic and real data sets.

1. Introduction

Today, in our daily lives, social networking enables us to contact our colleagues, friends, and families through applications, such as Facebook, Twitter, Linkedin, Google+, YouTube, and ResearchGate (Ferrag et al., 2017).Large amounts of data are generated by such communication on social networking, they will be explored for marketing, advertising, data mining and so on. The large amount of personal data that users share on social networks makes them a desirable target for attackers (Rathore et al., 2017). That suggests there is a potential risk for users' privacy being exposed. For example, users' identities, attributes, and relationships might be disclosed if the social graph released without being properly anonymized. Therefore, privacy protection has become one of the biggest problems with the progress of big data (Yu, 2016).

Graphs provide a powerful primitive for modeling data in a variety of applications. Nodes in graphs usually represent real world objects, and edges indicate relationships between objects (Yu et al., 2017) (Wang et al., 2018). Normally, data owners may release their data with users' identities hidden by Navïve anonymization. For example, the well-known Zachary karate club network (see Fig. 1), in which nodes represent the members of the club, and edges represent the relationships between members, is the navïvely anonymized social network. However, Navïve anonymization cannot protect users' privacy while adversaries own some background knowledge about users which can be modeled as attack graph(Enoch et al., 2019), i.e., the 1-neighborhood graphs (see Fig. 2(a) and (b)). Based on background knowledge, there exist re-identification or de-anonymization attacks (Narayanan and Shmatikov, 2009) (Ji et al., 2016) (Qian et al., 2017) (Ji et al., 2017) (Li et al., 2020) against graph structure, i.e., degree sequence attack (Kiabod et al., 2019), 1-neighborhood graph attack (Zhou and Pei, 2008), subgraph attack (Zou et al., 2009). Privacy risks can be broadly categorized into identity disclosure(Liu and Terzi, 2008),membership disclosure (Liu et al., 2016) and content disclosure (Cai et al., 2020).

To alleviate privacy problem in social networks, many privacy preserving mechanisms have been proposed, such as *k*-anonymity(Campan and Truta, 2008), differential privacy (Dwork et al., 2012), reputation (Cai et al., 2020), encryption (Ding et al., 2019). In *k*-anonymity, the probability of node re-identification is less than $1/k$, and the trade-off between data utility and privacy can be adjusted according to the value of k. Differential privacy provides a perturbation method to minimize the probability of identifying individual records by adding specific noise. It has been widely used to perturb statistical values of graph data, such as degree distribution, frequent subgraph mining, and triangle counting (Ding et al., 2021). Reputation-based and encryption methods are generally used for content disclosure in scenarios where users exchange information. Comparing with differential privacy,

[*] Corresponding author.
E-mail address: xuli@fjnu.edu.cn (L. Xu).

https://doi.org/10.1016/j.jnca.2021.103214
Received 1 March 2021; Received in revised form 17 August 2021; Accepted 5 September 2021
Available online 21 September 2021

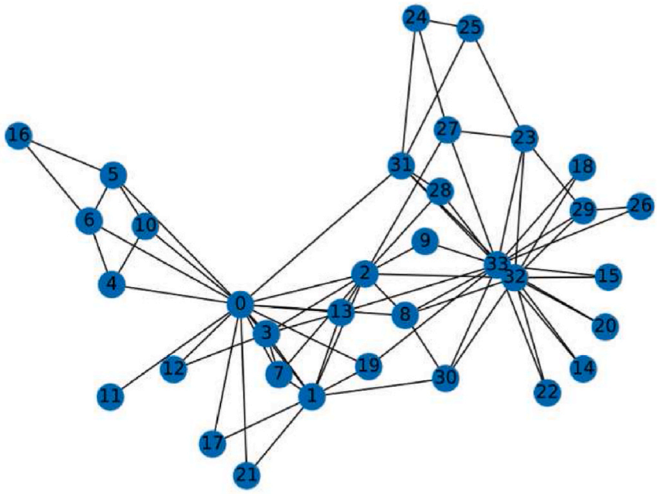


Fig. 1. Karate club graph.

k-anonymity is implemented by adding or deleting nodes and edges to make one node indistinguishable from the other *k* - 1 nodes. It can support not only statistical values query but also structural data query. Moreover, itis straightforward to demonstrate the security of *k*-anonymity, and is suitable for protecting privacy in social networks. In this paper, we consider the structure attack: 1-neighborhood graph attack, because it is more difficult for an attacker to collect the information beyond a one-hop neighborhood (Liu et al., 2017). The challenges is how to balance in trade-off between privacy and data utility. To address this issue, we propose a Graph Partition based Privacy-preserving Scheme, named GPPS, in social networks to achieve the *k*-anonymity against privacy disclosure while maintaining data utility. Our scheme consists of two main steps: 1) Node Clustering; 2) Graph Anonymization. First, inspired by literature (Li et al., 2016), we separate nodes into *T* clusters according to node similarity. This problem is addressed by RatioCut based spectral clustering algorithm (Von Luxburg, 2007), finding a partition of the similarity graph such that the edges between different clusters have a low weight and the edges within the same clusters have a high weight. The advantage is the size of each cluster is approximately equal. In our work, we modify the spectral clustering algorithm to improve the accuracy of node cluster, introducing the notion of degree-based entropy which can used to measure network heterogeneity (Cao et al., 2014) when computing node similarity. In graph anonymity step, to achieve *k*-anonymity, we use maximum weight bipartite graph matching method to calculate the cost of node matching and find the optimal solution in the graph modification in each cluster. We do not modify the 1-neighborhood graphs to be isomorphism, but node indistinguishable, the attacker cannot determine if the 1-neighborhood graph of the node in the anonymous graph is the same as in the original graph. Therefore, it can not only achieve *k*-anonymity but also maintain the utility of graph data.

The main contributions of this paper are summarized as follows:

1. We propose a graph partition based *k*-anonymous scheme to preserve the identity privacy of individuals in social networks, we convert the node clustering problem into graph partition problem, and in order to achieve balanced graph partition, we propose a modified RatioCut based spectral clustering algorithm.
2. We introduce the degree-based entropy to measure the network heterogeneity, and consider it as one of the metrics to calculate the similarity of node structure. This method helps to improve the accuracy of node clustering.
3. We develop a maximum weight bipartite graph matching algorithm based graph modification method on original graph to achieve *k*-anonymity and maintain the data utility. The experiment results illustrate that the proposed GPPS is effective and efficient both on synthetic and real datasets.

The rest of the paper is organized as follows. The notions, terminologies and the problem description are introduced in Section 3. The strategies are elaborated in Section 4. Section 5 gives the experimental analysis on our scheme respectively. The validation results are presented in this section as well. We conclude this paper in Section 6.

2. Related works

Data owners usually publish anonymized data for individual privacy protection. Navïve anonymous approaches which just remove the identities of individuals could not guarantee privacy. For example, attackers might use certain background knowledge to re-identify the nodes. To defend the re-identification attacks, a number of approaches have been proposed, which can be divided into three categories: nodes and edges perturbation, k-anonymity, and differential privacy (Ding et al., 2021).

Nodes and edges perturbation approaches are based on adding or deleting nodes and edges. Hay et al. (2007) proposed a method called random perturbation using Rand add/del to anonymize graphs, which randomly remove p edges, and then randomly add p edges. The main advantages of this method are that it is not only simple, but also of low complexity. However, the important nodes cannot be protected well, and can be re-identified. Ying and Wu(Ying and Wu, 2008) proposed an eigenvalues based random graph modification scheme which randomly deleting and swapping edges in the graph with less information loss. The approaches help to maintain the structural properties of social networks by maintaining a role based on the concept of rule equivalence in social networks, or edge intermediate based variation of limiting shortest paths. Yuan et al. (2013) defined a k-degree-l-diversity anonymity model that protects structural information as well as sensitive labels of individuals, the scheme is based on adding noise nodes into the original

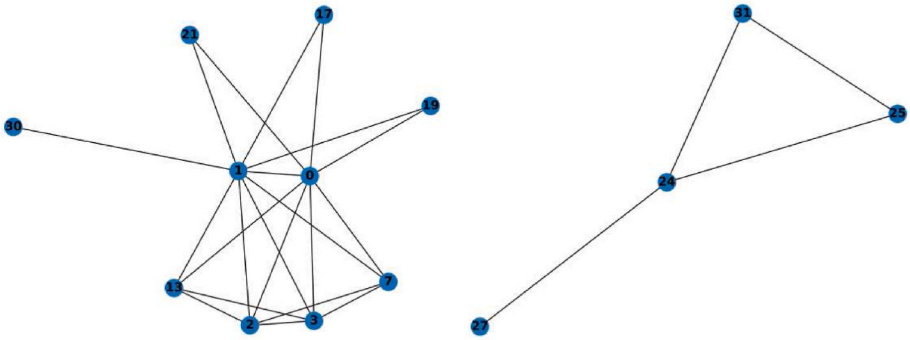


Fig. 2. 1-neighborhood graphs of node 1 and node 24 of the Karate club graph.

graph and editing edges with the consideration of guarantee the average path length. Liu et al. (2016) utilized a set of supervised machine learning techniques to predict the necessary random walk length based on the structural characteristics of a social graph, and generated a fake link to replace a real link between two users to protect link privacy.

K-anonymity has become the most widely used method to protect individuals' privacy in social network data publishing (Ding et al., 2021). Campan and Truta (2008) proposed a k-anonymity scheme, in which each node should be indistinguishable from at least k-1 nodes based on both structural information and nodes' attributes such that anonymized nodes cannot be re-identified with the probability larger than 1/k. Generally, k-anonymity approaches can be divided into k-Degree Anonymity (Liu and Terzi, 2008) which is used to defend degree attacks, k-Neighborhood Anonymity (Zhou and Pei, 2008) which is used to defend neighborhood graph attacks, k-Automorphism (Zou et al., 2009) which is used to defend subgraph attacks, and is based on the subgraph isomorphic technology (Ding et al., 2021). In order to achieve k-anonymity, graph editing or graph modification based approaches have been used. The aim of graph editing(or graph modification) problems is to modify a given graph by applying a bounded number of permitted operations in order to satisfy a certain property, including vertex deletions, edge deletions and edge additions, sometimes vertex additions are also permitted (Golovach and Mertzios, 2016). Liu and Terzi (2008) proposed a strategy to accomplish node degree based k-anonymity,in which each node in graph has at least other k-1 nodes of the same degree. Zhou and Pei (2008) proposed a scheme to against the 1-neighborhood attack. For each vertex v, there are k-1other nodes which have the isomorphic1-neighborhood graphs to v.Zou, Chen and Ozsu (Zou et al., 2009)proposed a k-automorphism scheme to preserve privacy. They anonymize the data graph through three steps graph partition, block alignment and edge copy to build k-automorphism anonymized graph. Cheng, Fu and Liu (Cheng et al., 2010) identified two realistic targets of attacks, NodeInfo and LinkInfo and then proposed a solution to form k pairwise isomorphic subgraphs so that the graph is k-isomorphic.The sensitive attributes of nodes are protected by anatomy model(Xiao and Tao, 2006) in a k-isomorphic graph. Li et al. (2016) proposed a graph partition based framework for privacy preserving graph data publication, which is designed to accommodate various datasets including social networks, temporal and spatial sequences. They defined graph-based privacy criterion and utility metrics to quantify the privacy and utility measurements of the anonymity data sets. Liu et al. (2017) defined weighted 1*-neighborhood attacks, which assume that attackers have some background knowledge about both individuals' 1-neighborhood graphs and related degrees and edge weights. In order to resist this kind of attacks, they proposed a heuristic indistinguishable group anonymous scheme to anonymize a weighted social graph. The 1*-neighborhood graphs of nodes in the same group are probabilistic indistinguishable and the published graph has high utility. Ding et al. (2021) measure the utility with a new information loss matrix, based on which a k-decomposition algorithm and a privacy preserving framework are developed for graph anonymization, and the proposed solution can be proved to achieve k-anonymity.

Dwork et al. (2012) designed differential privacy technique to solve the privacy protection problem in statistical databases, it can provide a mathematical security proof. It achieves privacy protection via injecting random noise into the query results. Kasiviswanathan et al. (2013) suggested projection operations on statistics to achieve low sensitivity and high data availability. Day et al. (2016) investigated the degree distribution publishing problem of a graph under node-DP by exploring the projection method to reduce the degree sensitivity, the proposed approaches are based on aggregation and cumulative histogram. Gao and Li (2019) proposed a novel anonymization scheme which preserves the persistent homology of the graph while satisfying the differential privacy. Based on the differential privacy model, Huang et al. (2020) proposed a privacy preserving approach, which combined clustering and randomization algorithms. Moreover, to objectively evaluate the privacy-preserving strength, they proposed a privacy measure algorithm against graph structure and degree attacks. Differential privacy differs from traditional methods and provides strong privacy guarantees without assuming the background knowledge of the attackers obtained.

Existing representative privacy-preserving approaches in social networks have corresponding features. Nodes and edges perturbation approaches are relatively simple and with higher data utility, however, the protection strength is not enough. The privacy protection strength of k-anonymity approaches depends on the value of k, the larger the k, the smaller the data utility. Therefore, the user can choose the appropriate k value as needed. Comparing with k-anonymity, differential privacy is usually used to protect various statistical values of graph data, such as degree distribution, edge weights, however, it is difficult to protect the structural information of graphs. Generally speaking, existing researches about privacy preserving in social networks can protect the privacy of users, however, k-anonymity is a better choice when suffering from a graph structure attack. However, privacy protection schemes require a trade-off between data utility and privacy (Ninggal and Abawajy, 2015). In our approach, we focus on 1-neighborhood graph attack, and based on graph partition, we can achieve k-anonymity meanwhile guarantee better data utility.

3. Preliminary

3.1. System model

In this paper, a social network is modeled as a simple undirected graph G = (V, E), where V is the node set which represent the individuals in the social networks, E is a set of edges which represent the relationships such as friendship, partnership between individuals. The cardinalities of V and E are denoted by $|\mathrm{V}|$ and $|\mathrm{E}|$ respectively, we assume that $|\mathrm{V}| = \mathrm{n}$, $|\mathrm{E}| = \mathrm{m}$. Zhou and Pei (2011) pointed out that it was difficult to obtain target's information beyond one-hop neighborhood, due to the small-world characteristic of social networks. Therefore, we assume that attackers have knowledge about the target's1-neighborhood graph.

Definition 1. (*1-Neighborhood Graph*) G(v) = <V(v), E(v)>, where V (v) is the set of neighborhood nodes of v and $\mathrm{V(v)} = \{\mathrm{u}|(\mathrm{u,v}) \in \mathrm{E}\} \cup \{\mathrm{v}\}$. E(v) is the set of edges between the nodes in V(v) and $\mathrm{E(v)} = \{(\mathrm{u,v})|\mathrm{u,v} \in \mathrm{V(v)} \wedge (\mathrm{u,v}) \in \mathrm{E}\}$.

To protect the individuals' identity privacy from re-identify attacks, the social network graph data will be anonymized to $\widetilde{\mathrm{G}} = (\widetilde{\mathrm{V}},\widetilde{\mathrm{E}})$, before being released. Inspired by (Liu et al., 2017), we modify the graph structure to achieve k-neighborhood anonymity by graph modification under the constrain of $|\mathrm{V}| = |\widetilde{\mathrm{V}}|$.

Definition 2. (*k-Neighborhood Anonymity*) Given a graph G(V, E), G satisfy k-neighborhood anonymity, if for each node $\mathrm{v} \in \mathrm{V}$, there are at least $\mathrm{k} -1$ other nodes have the same 1-neighborhood graphs with v.

Definition 3. (*Node Indistinguishability*) Node u and v are indistinguishable if an observer cannot decide whether or not $\mathrm{G(u)} \neq \mathrm{G(v)}$ in the original graph G, by comparing $\widetilde{\mathrm{G}}(\mathrm{u})$ and $\widetilde{\mathrm{G}}(\mathrm{v})$.

3.2. Problem statement

Given an undirected and unlabeled graph G, we try to obtain an anonymous graph $\widetilde{\mathrm{G}}$, such that no attacker can re-identify the nodes with the probability higher than 1/k and the information loss is minimized, which is formally stated as follows:

For original graph $G = (V, E)$ and anonymous graph $\widetilde{\mathrm{G}} = (\widetilde{\mathrm{V}}, \widetilde{\mathrm{E}})$, where $G(v)=(V(v),E(v))$ is the 1-neighborhood graph of node v, $\mathrm{G(v)} \subseteqq \mathrm{G}$, and $\widetilde{\mathrm{G}}(\mathrm{v}') = (\widetilde{\mathrm{V}}(\mathrm{v}'), \widetilde{\mathrm{E}}(\mathrm{v}'))$ is the 1-neighborhoodgraph of node v', $\widetilde{\mathrm{G}}(\mathrm{v}') \subseteqq \widetilde{\mathrm{G}}$. There exists a bijective function $\pi: V(v) \rightarrow V(u)$, such that for

Table 1
Meanings of symbols used.

Symbol	Meaning
G	Original graph
$\tilde{G}$	Anonymized graph
G^S	Similarity graph
G^B	Bipartitie graph
BC	Betweeness centrility
w_{ij}^S	*Similarity of node* v_i and v_j as edge weight in $\tilde{G}$
Lc	Local clustering coefficient
C_i	*i*th cluster
$I_f(G)$	Degree based graph Entropy
$f_{sim}()$	Similarity function
$Cost()$	cost of graphs modification

each $e = (v_i, v_j) \in G(v)$, there exists an edge $e' = (v_i', v_j') \in \widetilde{G}(v')$, ensuring $G(v) \cong \widetilde{G}(v')$, that is $Pr(G(v) \cong \widetilde{G}(v')) \leq 1/k$.

To be convenient, we summarize the commonly used symbols of this paper in Table 1.

4. The proposed strategies

In this paper, we propose a *k*-anonymity method based on RatioCut graph partition to protect the identity privacy of individuals in social networks. This method has a lower amount of information loss. It is a two step strategy to achieve *k*-anonymity. First, we use the RatioCut partition method to clustering the nodes into *K* clusters. Secondly, we modify the 1-neighborhood graphs of these nodes to make them probabilistic indistinguishable and satisfy *k*-anonymity. The flow chat can be shown as Fig. 3.

4.1. Node clustering

The goal of node clustering is to partition the nodes in graph *G* into *T* disjoint clusters $\{C_1, C_2, \ldots, C_T\}$, so that nodes within the same cluster are generally close to each other in terms of graph structure while distant otherwise (Fan et al., 2020) (Li et al., 2021).

Therefore, node clustering problem can be regarded to find a partition of the graph so that nodes in the same cluster are more similar to each other than nodes between different clusters, that is, the edges within a cluster have higher weights than edges between clusters. For a given graph $G = (V, E)$,we translate *G* to similarity graph G^S to achieve node clustering. Spectral clustering is one of the most commonly used clustering methods, and its performance is better than the traditional clustering methods (Von Luxburg, 2007). Spectral clustering has been applied successfully in a large number of fields, including bio-informatics (Yu et al., 2012), community detection (Qin et al., 2016) (Javed et al., 2018) and so on. We use approximating RatioCut based spectral clustering approach to partition graph G^S to ensure similar nodes in the same cluster and the clusters are balanced that is the number of nodes in each cluster is approximately the same (Von Luxburg, 2007). As shown in algorithm1, the processing is divided into three steps:

1. *Pairwise similarity computing*: we consider some metrics to compute the similarity between pairwise nodes.
2. *Similarity graph constructing*: we select the *K* nearest nodes as node's neighbors, so that, the similarity graph matrix is a sparse matrix.
3. *Similarity Graph partition*: we use the RatioCut graph partition scheme to partition the similarity graph into *T* subgraphs, so that the size of subgraphs are approximately equal.

Algorithm 1
Node Clustering.

```
Input: $G$
  Output: clusters $C_i$, $i = 1, 2, \ldots, T$
  for $i = 1$ to $n$ do
     construct1-neighborhoodgraph$G(v_i)$ of each node $v_i$
     compute $x_i = <D(v_i), BC(v_i), Lc(v_i), I_f(G(v_i))>$
  end for
  Similarity graph construction seeing Algorithm 2
  similarity graph partition seeing Algorithm 3
  obtain the clusters $C_i$, $i = 1, 2, \ldots, T$
  return $C_i$, $i = 1, 2, \ldots, T$
```

4.1.1. Pairwise similarity calculating

The effect of spectral clustering depends greatly on the similarity measurement used to construct the similarity matrix (Ye and Sakurai, 2016). In order to construct the similarity graph, we consider the following metrics to calculate the similarity between each pair of nodes: node degree, betweenness centrality, local clustering coefficient, and degree-based graph entropy.

Definition 4. (*Local Clustering Coefficent*) (Liu et al., 2017) $Lc(v_i) = \mu_G(v_i)/\omega_G(v_i)$, *where* $\mu_G(v_i)$ *and* $\omega_G(v_i)$ *are the numbers of triangles and triples in* $G(v_i)$, *respectively.*

Definition 5. (*Betweenness Centrality*) (Brandes, 2001) *BC*(*v*) *of node v in graph G is the fraction of the shortest paths between all pairs of nodes in the graph that pass through v,* $BC(v) = \sum_{s,t \in V, s \neq t \neq v} \frac{\sigma_{st}(v)}{\sigma(v)}$.

Definition 6. (*Degree Based Graph Entropy*) (Cao et al., 2014) *Let G* =

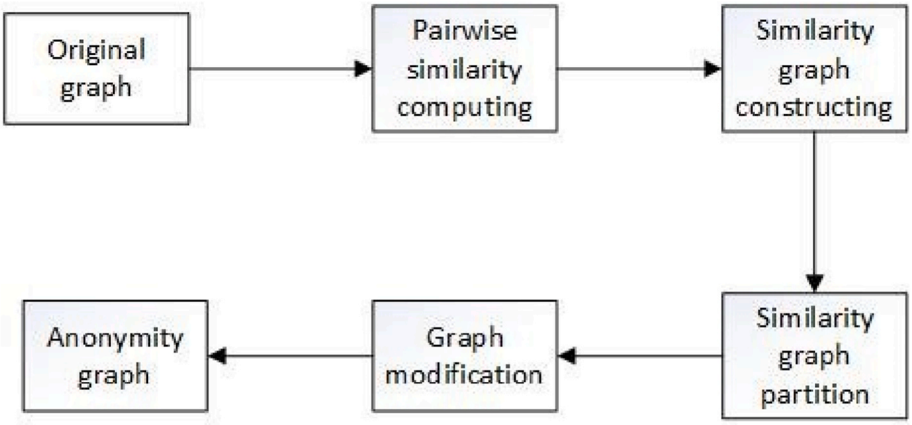

Original graph
Pairwise similarity computing
Similarity graph constructing
Similarity graph partition
Graph modification
Anonymity graph

Fig. 3. Flow chart of GPPS.

(V, E) be a connected graph. For a given v ∈ V *and anarbitrary real number* $\alpha \in$ R*, the degree-based graph entropy*

$$I_f(G) = -\sum_{i=1}^{n} \frac{d_i^\alpha}{\sum_{j=1}^{n} d_j^\alpha} \log\left(\frac{d_i^\alpha}{\sum_{j=1}^{n} d_j^\alpha}\right)$$

Let $\alpha = 1$, thus,

$$I_f(G) = \log\left(\sum_{i=1}^{n} d_i\right) - \sum_{i=1}^{n} \frac{d_i}{\sum_{j=1}^{n} d_j} \log d_i = \log(2m) - \frac{1}{2m}\sum_{i=1}^{n} d_i \log d_i$$

For every node $v_i \in G$, we compute $BC(v_i), Lc(v_i), I_f(G(v_i)$ respectively. We call the vector $x_i = < D(v_i), BC(v_i), Lc(v_i), I_f(G(v_i) >$ the node vector. Thus, the similarity function of every pair of nodes v_i, v_j can be defined as

$$f_{sim}\left(v_i, v_j\right) = e^{-\frac{\left|\left|x_i - x_j\right|\right|^2}{2\sigma^2}} \quad (1)$$

4.1.2. *Similarity graph constructing*

Algorithm 2
Similarity Graph Constructing.

```
Input: Node vector $x_i$, $i = 1, ..., n$
  Output: $G^S$
  for $i = 1$ to $n$ do
    for $j = 1$ to $n$ do
      Compute similarity $f_{sim}(v_i, v_j)$ between node $v_i$ and node $v_j$ according to equation(1)
    end for
  end for
  for $i = 1$ to $n$ do
    select the first $K$ nearest nodes as the neighbors of node $v_i$
    add edges between $v_i$ and it's neighbors and obtain similarity graph $G^S$
  end for
  return $G^S$
```

In this step, we construct a weighted similarity graph named K-nearest neighbor graph, whose neighborhood relationship is symmetric. Suppose $G^S = (V^S, E^S, W^S)$, where V^S represent the nodes, E^S represents the edges, W^S represent the weights on the edges. If v_j^S is one of the first K similar nodes to v_i^S, then $e_{ij}^S \in E^S$, and w_{ij}^S represents the similarity of v_i^S and $v_j^S w_{ij}^S = f_{sim}(v_i^S, v_j^S)$, if v_j^S is the K-nearest neighbors of v_i^S, otherwise $w_{ij}^S = 0$. Therefore, the degree of node v_i is defined as the sum of weights of the edges adjacent to v_i, $d_i = \sum_{t=i}^{N_i} d_t$, where N_i is the number of the neighbors of v_i. We use D to represent the degree matrix, $D = (d_{ij})_{n\times n}$, $d_{ij} = 1$, if $i = j$, else $d_{ij} = 0$. W is the adjacency matrix of G^S, where $w_{i,j}$ is the weight of the edge between node v_i and v_j. The detail is shown as Algorithm 2.

4.1.3. *Similarity graph partition*

In order to cluster the nodes in the original graph, first, we convert the original graph into a similarity graph. In this way, the node clustering problem is transformed into a problem of how to partition similarity graph into T subsets such that the cut is minimum, $cut(A_1, A_2, \ldots, A_T) = \frac{1}{2}\sum_{i}^{T} W(A_i, \overline{A_i})$. In order to make the subset balance, we replace the cut with RatioCut (Von Luxburg, 2007), $RatioCut(A_1, A_2, \ldots, A_T) = \sum_{i=1}^{T}\frac{cut(A_i, \overline{A_i})}{|A_j|}$. For a similarity graph G^S, W is the similarity matrix, its Laplacian matrix can be calculated as $L = D$-W. Suppose that $\lambda_1, \lambda_2, \ldots, \lambda_T$ are the smallest T eigenvalues of L, $min(RatioCut(A_1, A_2, \ldots, A_T) = \sum_{i=1}^{T}\lambda_i$. Then, compute the eigenvectors $x_1, x_2, \ldots, x_T$ corresponding to $\lambda_i, i = 1, 2, \ldots, T$.Then $y_i = \frac{x_i}{\sqrt{N_i}}$, $i = 1, 2, \ldots, T$, $Y = (y_i)_{T\times n}$. The minimum RatioCut problem can be relaxed as $\min tr(Y^T LY)$ subject to $Y^T Y = I$. Consider each column of matrix Y as one node, next, we utilize k-means algorithm to partition these n nodes into T clusters, $C_1, C_2, \ldots, C_T$. Here, in order to achieve k-anonymity, that is, the number of nodes in one cluster is in $[k, 2k)$. Thus, T takes value from $\left[\frac{n}{2k-1}, \frac{n}{k}\right)$. The detail is shown as Algorithm 3.

Algorithm 3
Similarity Graph Partition.

```
Input: $G^S$
  Output: clusters $C_1, C_2, \ldots, C_T$
  Construct adjacent matrix and degree matrix of graph $G^S$, denoted as $W$ and $D$, respectively
  Compute Laplacian matrix $L = D$ - $W$
  Compute the eigenvalue of $L$, the first $T$ eigenvalues are denoted as $\lambda_i, i = 1, 2, \ldots, T$
  Compute the correspondence eigenvectors of the first $T$ eigenvalues denoted as $x_1, x_2, \ldots, x_T$
  Let $y_i = \frac{x_i}{\sqrt{N_i}}, i = 1, 2, \ldots, T$, set $Y = (y_i)_{T\times n}$
  Consider one column of matrix $Y$ as one node
  Utilize $k$-means algorithm to partition these $n$ nodes into $T$ clusters
  return $C_1, C_2, \ldots, C_T$
```

4.2. *Graph anonymity*

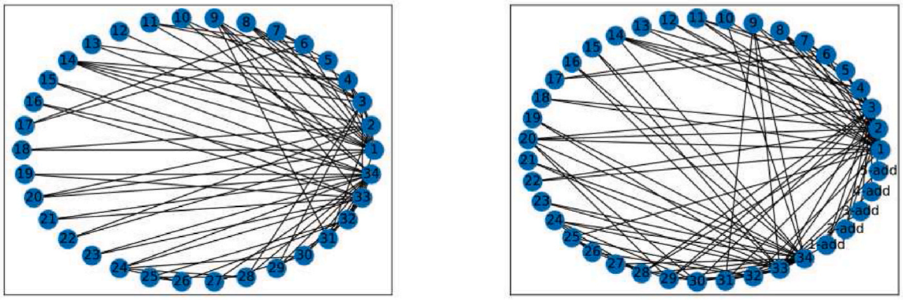
After node clustering, there are T clusters, where each cluster has about $[k, 2k)$ nodes. Next, we are going to modify the original graph to achieve k-anonymity. In order to guarantee the utility of the anonymous graph, we hope to reduce the cost of graph modification. First, for each pair of neighborhood graphs in the same cluster, we calculate the cost of modifying them to be isomorphic. Second, we choose the node that minimizes the total modification cost as the seed of each cluster, and modify the1-neighborhood graphs of others to make them isomorphic with the 1-neighborhood graph of the seed node. At last, we modify above mentioned 1-neighborhood graph according to the matching results to make anonymous graph achieve k-anonymous.

We use maximal bipartite graph matching algorithm (Sankowsk, 2009) to compute the cost of graph modifying. First, in arbitrary cluster C_i, sort the nodes with descending order of node degree. The size of nodes' neighborhood graphs in the same cluster might not be equal, therefore, before executing graph matching, we add the dummy nodes into graphs to ensure the size of graphs being equal in the same cluster, assume that the size of neighborhood graph in.

Algorithm 4
Graph Anonymity.

```
Input: $G, C_i, i = 1, 2, \ldots, T$
  Output: $\widetilde{G}$
  for $i = 1$ to $n$ do
    construct 1-neighborhood graph $G(v_i)$ of each node $v_i$
    sort all the clusters with descending order of the maximal node degree
  end for
  for $i = 1$ to $T$ do
    for $j = 1$ to $N_i$ do
      add dummy nodes into each 1-neighborhood graph to make the number of nodes equal to the maximum nodes in each cluster.
    end for
  end for
  for $i = 1$ to $T$ do
    for $j = 1$ to $N_i$ do
      for $t = 1$ to $N_i$ do
        Compute the matching cost of every pair of nodes $(v_j^i, v_t^i)$
```

(*continued on next page*)



(a) Original Karate

(b) 3-anonymousKarate

Fig. 4. Example of 3-anonymous Karate graph. (a) The original Karate graph; (b) An 3-anonymous Karate graph in which each node cannot be re-identify with the probability higher than 1/3. The nodes with label "add" are fake nodes and link them to corresponding nodes.

Algorithm 4 (*continued*)

```
            end for
        Compute $\sum_{t=1}^{N_i} cost(v_j^i, v_t^i)$
    end for
    Find a seed $v_s^i$ in each cluster $C_i$ to minimize cost $Cost(C_i)$
end for
utilize Algorithm 5 to modify the original graph to anonymize graph $\widetilde{G}$ which satisfy k-anonymity
return anonymous graph $\widetilde{G}$
```

The *i*th cluster is N_i. Then, for each pair of neighborhood graphs, execute the maximal graph matching algorithm to compute the matching cost and find a perfect matching between each pair of graphs. The matching cost of nodes in C_i between node v_j^i and node v_t^i is denoted as $cost(v_j^i, v_t^i)$. For an given $G(v_j)$, the total cost to modify other graphs into $G(v_j)$ can be defined as $Cost(C_j^i) = \sum_{t=1}^{N_i} cost(v_j^i, v_t^i)$, thus, the problem is formulized as

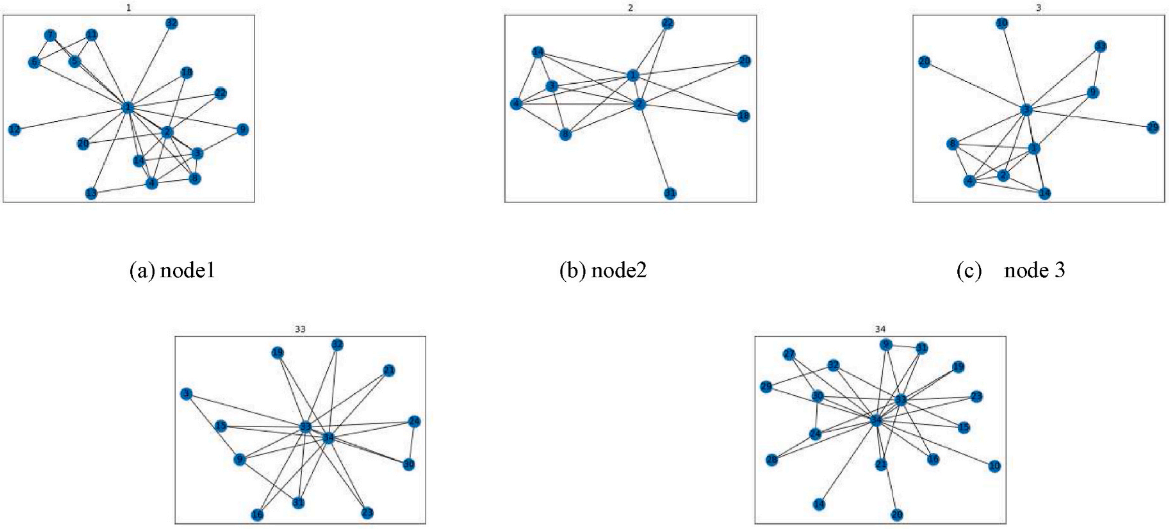
$$Cost\left(C^i\right) = \min_{j=1}^{N_i} Cost\left(C_j^i\right)$$

The solution is in each cluster C_i to find a node to minimize the total cost of graph modification, we consider this node as a seed node v_s^i, and save the maximal graph matching between $G(v_s^i)$ and other graphs. Next, we want to modify the graph G into graph $\widetilde{G}$ which satisfy k-anonymity, the processing is given as follow.

Algorithm 5
Graph Modification.

```
Input: $G, C_i, i = 1, 2, \ldots, T$, $G(v_j), j = 1, 2, \ldots, n$
Output: anonymous graph $\widetilde{G}$
for each cluster $C_i, i = 1, 2, \ldots, T$ do
    Sort nodes with descending order of node degree, obtain $v_j^{i'}, j = 1, 2, \ldots, N_i$
    for $j = 1$ to $N_i$ do
        if $Cost(v_j^i, v_s^i) > \delta$ then
            for matched node pair $v_{jw}^i$ in $G(v_j^i)$ and nodes $v_{jw}^s$ in $G(v_s^i)$
                if $degree(v_{jw}^i) < degree(v_{jw}^s)$ then
                    Find nodes $v_{jt}^i \in G(v_j^i)$ and $v_{jt}^i$ is not the neighbor of $v_{jw}^i$ which degrees are expected to be increased
                    add an edge between $v_{jt}^i$ and $v_{jw}^i$
                end if
                if $degree(v_{jw}^i) > degree(v_{jw}^s)$
```

(*continued on next page*)



(a) node1

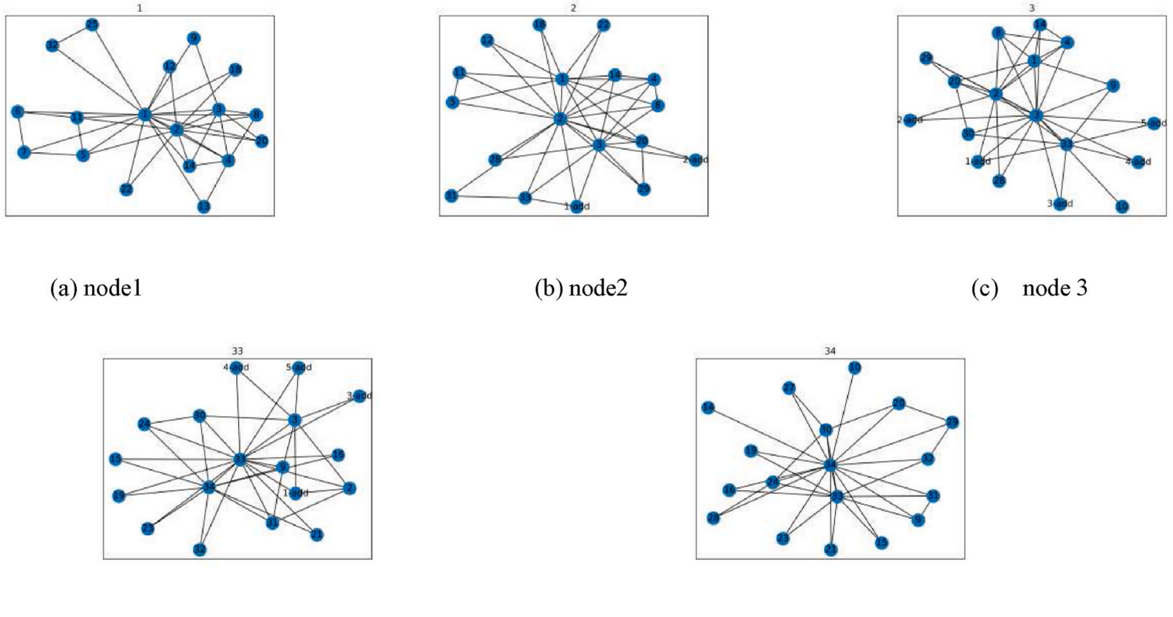
(b) node2

(c) node 3

(d) node 33

(e) node 34

Fig. 5. 1-neighborhood graphs of node 1, 2, 3, 33, 34 in Karate graph.



(a) node1

(b) node2

(c) node 3

(d) node 33

(e) node 34

Fig. 6. Modified 1-neighborhood graphs of node 1, 2, 3, 33, 34. The nodes in the same clusters must be node indistinguishability, add some nodes with label "add" and edges between "add" nodes and corresponding nodes such that node1, 2, 3, 33, 34 are node indistinguishability.

Algorithm 5 (*continued*)

```
                Find nodes $v_{jt}^i \in G(v_j^i)$ and $v_{jt}^i$ is the neighbor of $v_{jw}^i$ which degrees are expected to be decreased
                Choose the edge which has the smallest BC to delete
            end if
            Repeat until $degree(v_{jw}^i) = degree(v_{jw}^s)$
        end for
    end if
  end for
end for
return anonymous graph $\widetilde{G}$
```

For every two nodes v_j^i and v_t^i in the *ith* cluster, we create a weighted bipartite graph $G_B = (V_1, V_2, E_B, W)$, setting $V_1 = V(G(v_j^i))$, $V_2 = V(G(v_t^i))$, $E_B = V_1 \times V_2$ and W is the matching cost matrix. Let $u \in G(V_j^i)$, $u \in G(V_j^i)$, $w_{uv} = d_u - d_v$. Because $|V_1| = |V_2|$, we can find a minimal perfect matching M in G_B which is a set of edges in E such that each node is associated with only one edge and matches all the nodes. We apply the maximum weight bipartite matching algorithm (Sankowsk, 2009) to obtain the minimal perfect matching in polynomial time. Therefore, the cost of matching the bipartite graph is the sum of the costs of matching all nodes which can be calculated by $cost(G(v_j^i), G(v_t^i)) = \min\left(\sum_{u,v\notin V_D} w_{uv} + (\mu|V_D|)\right)$ where V_D is the set of dummy nodes, μ is the fixed cost value of adding a dummy node. When obtain the matching cost of each pair of neighborhood graphs, we can compute the minimal cluster cost and meanwhile find the matching seed. Assume that, v_s^i is the seed node in cluster C_i, for arbitrary node $v_j^i \in C_i, j \neq s$, if $\left|cost(G(v_j^i), G(v_s^i))\right| < \delta$, we consider they are similar, thus, we need not to modify $G(v_j^i)$.Otherwise, for node $v_{jw}^i \in G(v_j^i)$, if the degree of v_{jw}^i is smaller than v_s^i, that is $d(v_{jw}^i) < d(v_s^i)$, node v_{jw}^i needs to increase its degree. First, select nodes which should increase the degree as a candidate set, find a node $v_{jt}^i \in G(v_j^i)$ which need to increase its degree and they are not neighbors of each other. The processing continues until $d(v_{jw}^i) = d(v_s^i)$.If $d(v_{jw}^i) > d(v_s^i)$, we must delete the edges which adjacent to v_{jw}^i, find a node $v_{jt}^i \in G(v_j^s)$ which needs to decrease its degree, and the BC of the edge (v_{jw}^i, v_{jt}^i) is the smallest among the edges adjacent to v_{jw}^i.The detail is presented as Algorithm 4.

We then give the example of the proposed GPPS in Fig. 4, Fig. 4(a) is the original Karate club graphand Fig. 4(b) is the modified graph which satisfy 3-anonymous, Fig. 5 shows the original 1-neighborhood graphs, while Fig. 6 shows the modified 1-neighborhood graphs. In our scheme, nodes 1, 2, 3, 33, 34 are clustered into the same group, we modify their original 1-neighborhood graphs so that the probability of node indistinguishability can achieve 95%.

5. Theoretical analysis

5.1. Privacy analysis

Theorem 1. *From the anonymous graph $\widetilde{G}$, the probability of re-identifying the target cannot be higher than $1/k$, even if an attacker with the knowledge of any target's 1-neighborhood graph.*

Proof. The attacker has the knowledge about target's 1-neighborhood graph, he would like tore-identify a target from the published graph $\widetilde{G}$.

There are two possible consequences after searching $\widetilde{G}$:

Case 1. The attacker can find an exact match of the target.

Case 2. The attacker cannot find an exact match of the target.

For the first case, the attacker wants to re-identify node u, the 1-neighborhood graph of u in the original graph and published graph are denoted as $G(u)$, $\widetilde{G}(u)$, respectively. In this case, the attacker's knowledge is $G(u)$, and $G(u) = \widetilde{G}(u)$. In fact, we know that node u belongs to one cluster which has members more than k, in other words, there are at least $k - 1$ nodes have the same 1-neighborhood graph with node u, therefore, the attacker cannot re-identify the target with probability higher than $1/k$.

For the second case, the attacker knows $G(u)$, however, he cannot find an exact match of the $G(u)$. First, he might find a most similar match of $G(u)$, in this step, there might exist some deviation, then, the remaining part of Case 2 can be proven in a similar manner with Case 1. Therefore, in this case, the attacker cannot re-identify the target with

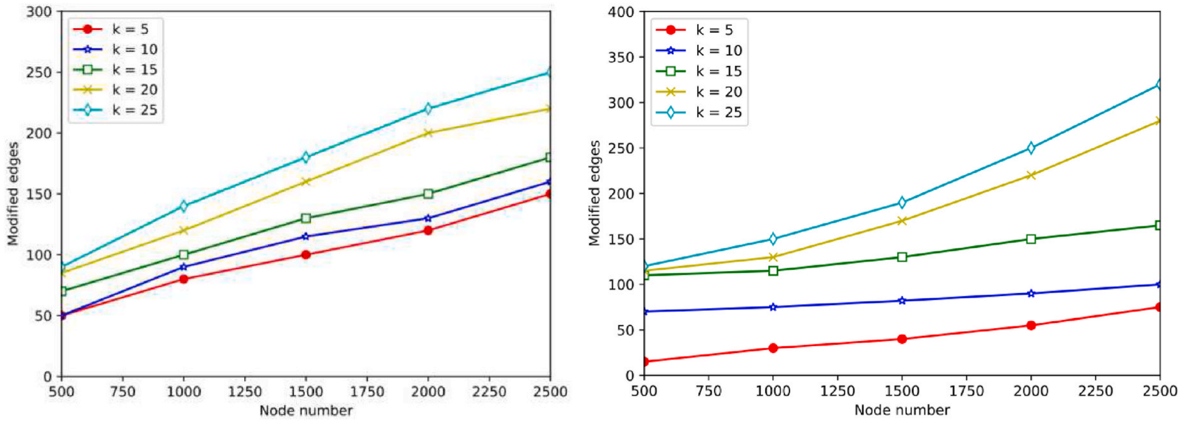


Fig. 7. Number of modified edges of synthetic data sets, the number of nodes increases from 500 to 2500.(a) The average degree(AVD) is 5; (b) The average degree (AVD) is 10.

probability higher than $1/k$.

In summary, the attacker cannot re-identify the target with probability higher than $1/k$.

5.2. Information loss

In our scheme, the processing of anonymizing graph contains node adding, edge swapping, deleting and adding, these operations lead to some information loss. To generate an anonymous graph $\widetilde{G}$ for the original graph G, our scheme first group the nodes into some clusters. In each cluster Ci, we modify the original 1-neighborhood graphs of nodes to make them indistinguishable. $MaxD_i$ denotes the largest degree in cluster C_i. Let n_i, m_i denote the number of nodes and edges in C_i. Let s^i_{wt}, a^i_{wt}, d^i_{wt} denote the number of swapped, added, deleted edges between $G(u_w)$ and $G(u_t)$ in cluster C_i, respectively. Therefore, the information loss IL_i for transferring C_i to $C_i^{'}$ can be calculated as $IL_i = NL_i + EL_i$, where $NL_i = \sum_{t=1}^{n_i}(MaxD_i - degree(v_t))$ and $EL_i = (\sum_{w,t=1}^{n_i} s^i_{wt} + \sum_{w,t=1}^{n_i} a^i_{wt} + \sum_{w,t=1}^{n_i} d^i_{wt})/m_i$. Therefore, the total information loss(IL) for anonymizing G to $\widetilde{G}$ can be calculated as $IL = \sum_{i=1}^{T} IL_i$.

6. Validation experiment

In this section, we will validate the performance of the proposed GPPS on both synthetic and real data sets. All the experiments were conducted in Python on a server running the Ubuntu 20.04.1 LTS operating system, multi-core of double Intel Xeon E5-2650, 2.20 GHz cpu and 755.6 GB RAM.

To explore the utility of the anonymized graph $\widetilde{G}$ of the proposed GPPS, we test the following three metrics:

1. *Average degree(AVD)*: The AVD of G can be calculated as $\sum_{v\in V} d_v/|V|$;
2. *Average clustering coefficient(ACC)*: The ACC of G can be calculated as $\sum_{v\in V} C_v/|V|$, where C_v is the local clustering coefficient of v;
3. *Average shortest path length(APL)*: We calculate shortest path length between each pair of nodes u and v in G and $\widetilde{G}$, denoted as *path(u, v)*, $\text{path}(u, \ v)^{'}$, respectively. Therefore, we can calculate APL and $APL^{'}$ in G and $\widetilde{G}$, $APL = 2\sum_{u,v\in V} path(u,v)/n(n-1)$.

6.1. Synthetic data sets

In our experiments, we generate synthetic data sets (Stanford Large Network Dataset Collection), the generated social networks have 500–2500 nodes the average node degree are 5 and 10, respectively. Figs. 7 and 8 show the number of modified edges and information loss considering different privacy requirements k and the number of nodes while the proposed strategy is employed. As the number of nodes increases, the number of modified edges increases and the information loss decreases. In addition, a larger k results in a larger number of modified edges as well as a greater information loss. This is because the uneven

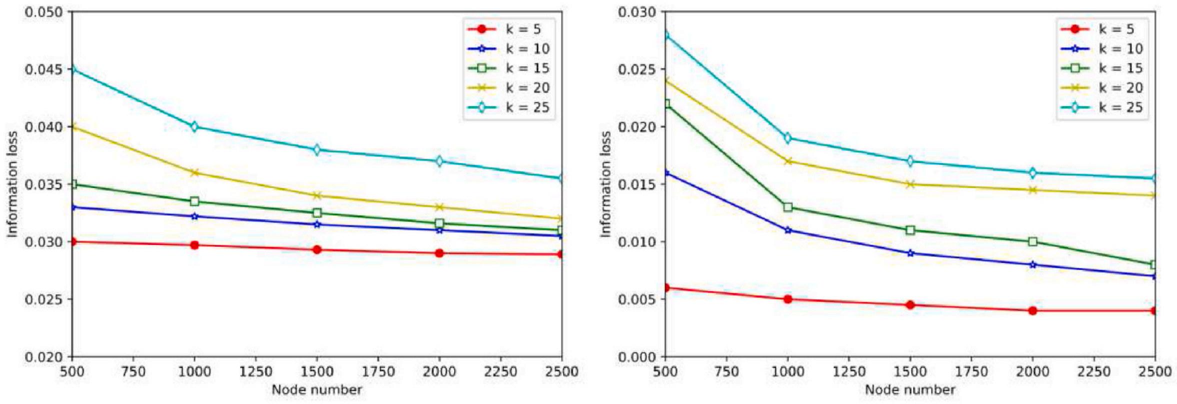


Fig. 8. Information loss of synthetic data sets with the number of nodes increases from 500 to 2500.(a) The average degree(AVD) is 5; (b) The average degree(AVD) is 10.

Table 2
Details of social networks.

Dataset	Nodes	Edges	AVD	ACC	APL
Facebook	4039	88234	44	0.605	4.7
HepTh	9877	25998	5.3	0.4714	7.4
Enron	23133	183831	10	0.497	6.4

degree distribution will result in a greater number of modified edges such that the information loss is even worse, while a larger k is adopted.

6.2. Real data sets

6.2.1. DataSets

There are lots of datasets available in the research of privacy protection in social networks. They come from different domains, such as online social networks, citation networks, collaboration networks, communication networks, and location-based social networks. These datasets are modeled as a variety of graphs, such as undirected graphs, directed graphs, weighted graphs, and node-labeled graphs. The choice of social networks in our experiment is mainly based on four reasons. First, the networks must be an undirected graph model consistent with our network model; Second, these networks must have different structural characteristics, because our solution depends on the network's structural characteristics; Third, the network must be large enough to represent a real social network; Fourth, choosing different networks helps to reduce the deviation of the performance measurement due to the characteristics of the specific network. For the above reasons, we selected three real data sets from the Stanford Network Analysis Project (SNAP) (Stanford Large Network Dataset Collection): Facebook, HepTh, and Enron. These datasets come from three domains, including: social networks, citation networks, and email networks.

The Facebook dataset consists of 4039 nodes and 88234 edges, which with smallest number of nodes but with largest average degrees(44) and largest average clustering coefficient(0.605). These factors may make Facebook more sensitive to edges and nodes perturbation. HepTH (High Energy Physics-Theory) with medium size consists of 9877 nodes and 25998 edges. HepTH collaboration network is from the e-print arXiv and covers scientific collaborations between authors' papers submitted to High Energy Physics-Theory category. If an author i co-authored a paper with author j, the graph contains an undirected edge from i to j. Enron email communication network consists of 23133 nodes and 183831 edges which covers all the email communication within a dataset of around half million emails. This data was originally made public, and posted to the web, by the Federal Energy Regulatory Commission during its investigation. Nodes of the network are email addresses and if an address i sent at least one email to address j, the graph contains an undirected edge from i to j. The details of these datasets are shown in Table 2.

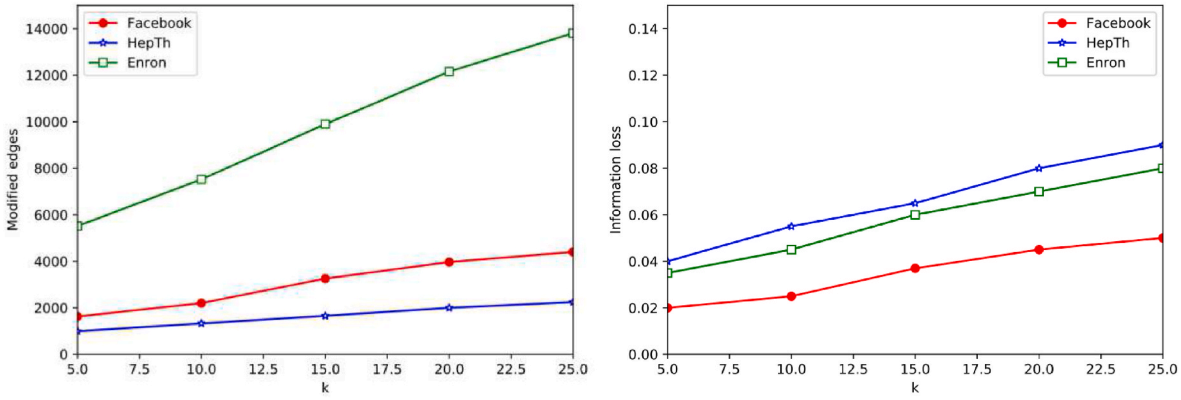
6.2.2. Utility

First, evaluate the modified edges and information loss, then consider the impact on the AVD, ACC, APL under different privacy requirements k. Besides these, it is also crucial to ensure the anonymized data is useful for data mining. We consider the impact on top influential nodes(TIN) in these data sets to find a set of users with the maximum influence in network. Observed from Fig. 9(a) and (b), we know that either the number of modified edges or information loss increase with privacy requirements k. It is obviously that the largest number of modified edges is required by Enron, while the greatest information loss is imposed on HepTh due to the same reason as that in Figs. 7 and 8.

Fig. 10(a)–(c) show the impact of privacy requirements k on percentage of either.

AVD, ACC or APL. It is clear that the percentage of either AVD, ACC or APL rises with the growth of k for Facebook, Enron and HepTH. The Facebook not only has the highest variation percentage of AVD but that of ACC and APL as well. The reason behind that is as follows. The severer the uneven degree distribution is, the larger variation of AVD will get and a greater variation the ACC will encounter. Besides, a larger ACC will result in a larger APL variation. Fig. 10(d) shows the percentage of influence maximization nodes remaining in different datasets, we can see that as k increases, our scheme works stable in all datasets. Our scheme can keep the maximum influence nodes almost greater than 95%, only in HepTh the percentage is slightly smaller than 95%. The effectiveness of our scheme in terms of AVD, ACC and APL on the Facebook is worse than the other two datasets, because our scheme is more sensitive to the distribution of degrees and the clustering coefficient.

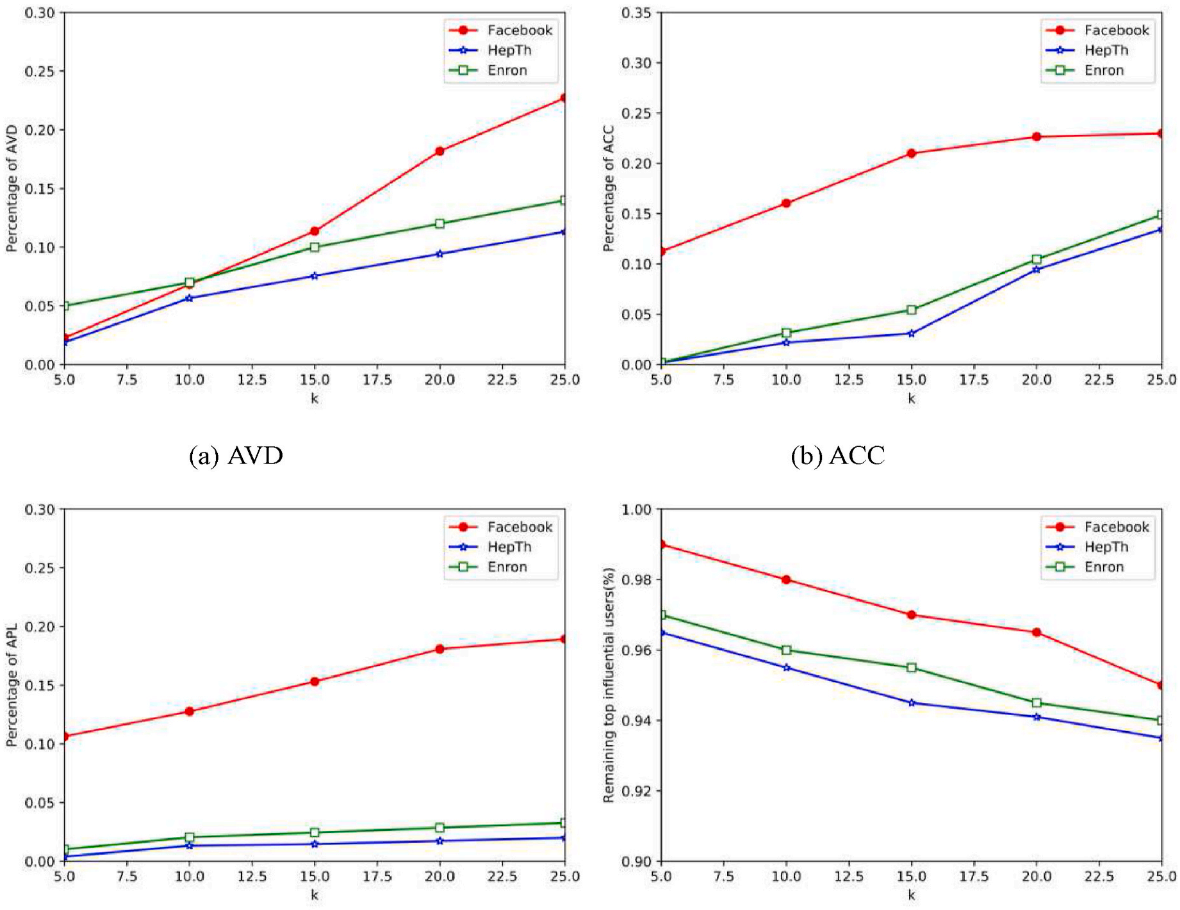
In order to show the effectiveness of our GPPS, we compare it with existing k-neighbor anonymity scheme. We compare the results with HIGA (Liu et al., 2017) on Facebook and Enron in terms of AVD, ACC, APL. Fig. 11 presents the results of evaluating GPPS and HIGA on Facebook. As it can be seen from Fig. 11(a), when k is small, the change of AVD in GPPS is less than HIGA, when k increases to 25, GPPS is slightly larger than HIGA. From Fig. 11(b), we can see the change of ACC is larger as k increases, it becomes less than HIGA. However, the change of APL in GPPS is about 2% smaller than that in HIGA. Fig. 12 presents the results of evaluating GPPS and HIGA on Enron. We can see that when k is small, HIGA's performance is better than GPPS, but as k increases, our GPPS is better than HIGA.



(a) Number of modified edges

(b) Information loss

Fig. 9. Number of modified edges & Information loss. (a) Number of modified edges increases as k increases, in Enron, the amount of change is largest; (b)Information loss increases as k increases, however, in each dataset, information loss is less than 10%.



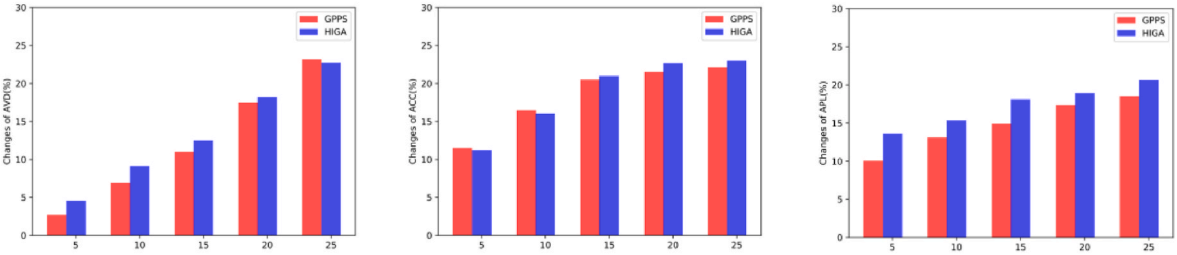
(a) AVD

(b) ACC

(c) APL

(d) TIN

Fig. 10. Utility of GPPS. (a) Change of AVD increases, as k increases, and the amount of change on Facebook is greater than the other two datasets; (b) Change of ACC increases as k increases, and the amount of change on Facebook is always greater than the other two datasets; (c) Change of APL increases as k increases, the amount of change on HepTh and Enron is small, however, the amount of change on Facebook is almost five times that of the other two data sets; (d) In the three data sets, the retention value of the most influential node is almost all greater than 95%, only ENRON is slightly lower than 95% when $k = 25$.



(a) AVD

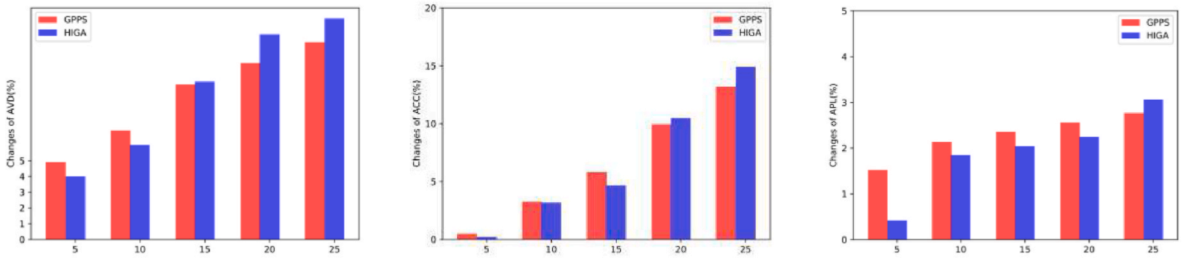
(b) ACC

(c) APL

Fig. 11. Utility comparison of GPPS and HIGA in Facebook for different k.(a) The change of average degree(AVD); (b) The change of average clustering coefficient (ACC); (c) The change of average path length(APL).

7. Conclusions

There has been an increasing interest in privacy disclosure problem due to more and more users release personal data to social platforms. These data contain users' private information, which make them subject to malicious attacks against users' privacy. Although graph anonymization can reduce the risk of privacy disclosure, malicious attackers might launch 1-neighborhood graph attack to obtain targets' identities. In this paper, we propose a Graph Partition based Privacy-preserving Scheme, named GPPS, in Social Networks to realize social graph anonymization. We utilize graph partition based k-anonymity to protect the identity privacy of individuals. In graph partition, the degree-based graph entropy is introduced to compute the similarity matrix in order to improve node clustering accuracy. Then, to achieve node indistinguishability, the graph modification is implemented, in which the graph information loss is minimized. The experiment results illustrate the security of privacy-preserving, utility and efficiency of our GPPS on both synthetic and real data sets. In our future work, we will try to introduce



(a) AVD (b) ACC (c) APL

Fig. 12. Comparing utilities of GPPS and HIGA in Enron for different *k*. (a) The change of average degree (AVD); (b) The change of average clustering coefficient (ACC); (c) The change of average path length (APL).

other privacy-preserving methods to defend subgraph attacks in social networks, e.g. uncertain graph method which converts an original graph into a weighted graph, where the weights on edges represent the probability of edges exist.

Credit roles

Hongyan Zhang: Conceptualization, Methodology, Writing – original draft. Limei Lin: Data curation, Software. Li Xu: Supervision, Writing – review & editing. Xiaoding Wang: Visualization, Formal analysis.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

National Natural Science Foundation of China (Nos. U1905211, 61771140, 61702100, 61702103), Fok Ying Tung Education Foundation (No. 171061), Natural Science Foundation of Fujian Province (No. 2020J01167), Educational Research Project for Young and Middle aged Teachers in Fujian Province (Science and Technology) (No. JAT200968).

References

Brandes, Ulrik, 2001. A faster algorithm for betweenness centrality. J. Math. Sociol. 25 (2), 163–177.

Cai, Y., Zhag, S., Xia, H., Fan, Y., Zhang, H., 2020. A privacy-preserving scheme for interactive messaging over online social networks. IEEE Internet Things J. 7 (8), 6817–6827.

Campan, A., Truta, T., 2008. A clustering approach for data and structural anonymity in social networks. In: Proc. Of the 2nd ACM SIGKDD Int. Workshop Privacy Security Trust in KDD, pp. 33–54.

Cao, S., Dehmer, M., Shi, Y., 2014. Extremality of degree-based graph entropies. Inf. Sci. 278, 22–33.

Cheng, J., Fu, A., Liu, J., K-isomorphism, 2010. Privacy preserving network publication against structural attacks. In: Proceedings of ACM SIGMOD International Conference on Management of Data, pp. 459–470.

Day, W., Li, N., Lyu, M., 2016. Publishing graph degree distribution with node differential privacy. In: Proceedings of the 2016 International Conference on Management of Data, vols. 123–138.

Ding, X., Liu, P., Jin, H., 2019. Privacy-preserving multi-keyword top-*k* similarity search over encrypted data. IEEE Trans. Dependable Secure Comput. 16 (2), 344–357.

Ding, X., Wang, C., Choo, K.K.R., Jin, H., 2021. A novel privacy preserving framework for large scale graph data publishing. IEEE Trans. Knowl. Data Eng. 1–13.

Dwork, C., Mcsherry, F., Nissin, K., Smith, A., 2012. Calibrating noise to sensitivity in private data analysis. Lect. Notes Comput. Sci. 3876 (8), 265-28.

Enoch, S., Hong, J., Kim, D., 2019. Security modeling and assessment of modern networks using time independent Graphical Security Models. J. Netw. Comput. Appl. 148, 102448. https://doi.org/10.1016/j.jnca.2019.102448.

Fan, S., Wang, X., Shi, C., Lu, E., Lin, K., Wang, B., 2020. One2multi graph autoencoder for multi-view graph clustering. In: Proc. Of the Web Conference, pp. 3070–3076.

Ferrag, M., Maglaras, L., Ahmim, A., 2017. Privacy-preserving schemes for ad hoc social networks: a survey. IEEE Commun. Surv. Tutor. 19 (4), 3015–3045.

Gao, T., Li, F., 2019. Preserving persistent homology in differentially private graph publications. In: Proc. of IEEE INFOCOM, pp. 2242–2250.

Golovach, P., Mertzios, G., 2016. Graph editing to a given degree sequence. Theor. Comput. Sci. 665, 1–12.

Hay, M., Miklau, G., Jensen, D., Weis, P., Srivastava, S., 2007. Anonymizing social networks. Int. J. Very Large Data Bases.

Huang, H., Zhang, D., Xiao, F., Wang, K., Gu, J., Wang, R., 2020. Privacy-preserving approach PBCN in social network with differential privacy. IEEE Trans. Netw. Serv. Manag. 17 (2), 931–945.

Javed, M.A., Younis, M.S., Latif, S., Qadir, J., Baig, A., 2018. Community detection in networks: a multi-disciplinary review. J. Netw. Comput. Appl. 108, 87–111.

Ji, S., Li, W., Srivatsa, M., Beyah, R., 2016. Structural data de-anonymization theory and practice. IEEE/ACM Trans. Netw. 24 (6), 1–14.

Ji, S., Mittal, P., Beyah, R., 2017. Graph data anonymization, de-anonymization attacks, and de-anonymizability quantification: a survey. IEEE Commun. Surv. Tutor. 19 (2), 1305–1326.

Kasiviswanathan, S.P., Nissim, K., Raskhodnikova, S., Smith, A., 2013. Analyzing graphs with node differential privacy. In: Proc. Of 10th Theory of Cryptography Conference on Theory of Cryptography, pp. 457–476.

Kiabod, M., Dehkordi, M., Barekatain, B., 2019. TSRAM: a time-saving k-degree anonymization method in social network. Expert Syst. Appl. 125, 378–396.

Li, X.-Y., Zhang, C., Jung, T., Qian, J., Chen, L., 2016. Graph-based privacy-preserving data publication. In: Proceedings of IEEE INFOCOM, pp. 1–9.

Li, H., Chen, Q., Zhu, H., Ma, D., Wen, H., Shen, X.S., 2020. Privacy leakage via de-anonymization and aggregation in heterogeneous social networks. Trans. Dependable Secur. Comput. 17 (2), 350–362. https://doi.org/10.1109/TDSC.2017.2754249.

Li, B., Pi, D., Lin, Y., Cui, Lin, 2021. DNC: a deep neural network-based clustering-oriented network embedding algorithm. J. Netw. Comput. Appl. 173, 102854.

Liu, K., Terzi, E., 2008. Towards identity anonymization on graphs. In: Proceedings of the 2008 ACM SIGMOD International Conference on Management of Data.

Liu, Y., Ji, S., Smartwalk, P. Mittal, 2016. Enhancing social network security via adaptive random walks. In: Proc. ACM SIGSAC Conf. Comput. Commun. Security, pp. 492–503.

Liu, Q., Wang, G., Li, F., Yang, S., Wu, J., 2017. Preserving privacy with probabilistic indistinguishability in weighted social networks. IEEE Trans. Parallel Distr. Syst. 28 (5), 1417–1429.

Narayanan, A., Shmatikov, V., 2009. De-anonymizing social networks. In: Proceedings of the 30th IEEE Symposium on Security and Privacy.

Ninggal, M., Abawajy, J., 2015. Utility-aware social network graph anonymization. J. Netw. Comput. Appl. 56, 137–148. https://doi.org/10.1016/j.jnca.2015.05.013.

Qian, J., Li, X., Zhang, C., Chen, L., Jung, T., Han, J., 2017. Social network de-anonymization and private inference with knowledge graph model. IEEE Trans. Dependable Secur. Comput. 99, 1–14.

Qin, X., Dai, W., Jiao, P., Wang, W., Yuan, N., 2016. A multi-similarity spectral clustering method for community detection in dynamic networks. Sci. Rep. 1–11. https://doi.org/10.1038/srep31454.

Rathore, S., Sharma, P.K., Loia, V., Jeong, Y.S., Park, J.H., 2017. Social network security: issues, challenges, threats, and solutions. Inf. Sci. 421, 43–69.

Sankowsk, P., 2009. Maximum weight bipartite matching in matrix multiplication time. Theor. Comput. Sci. 410, 4480–4488.

Stanford large network dataset collection, http://snap.stanford.edu/data.

Von Luxburg, U., 2007. A tutorial on spectral clustering. Stat. Comput. 17 (4), 395–416.

Wang, Q., Zhang, Y., Lu, X., Wang, Z., Qin, Z., Ren, K., 2018. Real-time and spatio-temporal crowd sourced social network data publishing with differential privacy. IEEE Trans. Dependable Secur. Comput. 15 (4), 591–606.

Xiao, X., Tao, Y., 2006. Anatomy: simple and effective privacy preservation. In: Proc. Of the 32nd International Conference on Very Large Data Bases. VLDB Endowment, pp. 139–150.

Ye, X., Sakurai, T., 2016. Robust similarity measure for spectral clustering based on shared neighbors. ETRI J. 38 (3), 540–550.

Ying, X., Wu, X., 2008. Randomizing social networks: a spectrum preserving approach. In: Proc. Of the 2008 SIAM International Conference on Data Mining, pp. 739–750.

Yu, Shui, 2016. Big privacy: challenges and opportunities of privacy study in the age of big data. IEEE Access 4, 2751–2763.

Yu, Z., Li, L., You, J., Wong, H., Han, G., 2012. Triple spectral clustering based consensus clustering framework for class discovery from cancer gene expression profiles. IEEE ACM Trans. Comput. Biol. Bioinf 9 (6), 751–1765.

Yu, S., Liu, M., Dou, W., Liu, X., Zhou, S., 2017. Networking for big data: a survey. IEEE Commun. Surv. Tutor. 19 (1), 531–549.

Yuan, M., Chen, L., Yu, P.S., Yu, T., 2013. Protecting sensitive lables in social network data anonymization. IEEE Trans. Knowl. Data Eng. 25 (3), 633–647.

Zhou, B., Pei, J., 2008. Preserving privacy in social networks against neighborhood attacks. In: Proceedings of 24th International Conference on Data Engineering, vols. 506–515. IEEE.

Zhou, B., Pei, J., 2011. The k-anonymity and l-diversity approaches for privacy preservation in social networks against neighborhood attacks. Knowl. Inf. Syst. 28 (1), 47–77.

Zou, L., Chen, L., Ozsu, M., K-automorphism, 2009. A general framework for privacy preserving network publication. Proc. VLDB Endow. 2 (1), 946–957.

Hongyan Zhang received the B.S. degree from Shandong University of Science and Technology in 2003, and the M.S. degree from Fujian Normal University in 2008. She is currently a Ph.D. candidate at Fujian Normal University, China. Her research interests include network security and wireless networks and communication.

Limei Lin received the M.S. and Ph.D. degrees in mathematics from the College of Mathematics and Informatics, Fujian Normal University, Fuzhou, China, in 2013 and 2016, respectively. She was a Visiting Scholar with Montclair State University in 2015. She was with Fujian Agriculture and Forestry University from 2016 to 2018. She is currently an Associate Professor with the College of Mathematics and Informatics, Fujian Normal University. She has authored or coauthored more than 30 papers in these areas, especially more than ten papers in IEEE journals, including IEEE TRANSACTIONS ON COMPUTERS, IEEE TRANSACTIONS ONPARALLEL AND DISTRIBUTED SYSTEMS, IEEE TRANSACTIONS ON DEPENDABLE AND SECURE COMPUTING, and IEEE TRANSACTIONS ON RELIABILITY. Her current research interests include interconnection networks, network security, and network reliability.

Li Xu received the B.S. and M.S. degrees from Fujian Normal University, Fuzhou, China, in 1992 and 2001, respectively, and the Ph.D. degree from Nanjing University of Posts and Telecommunications, Nanjing, China, in 2004. He is currently a Professor and Doctoral Supervisor with the College of Mathematics and Informatics, Fujian Normal University. He is currently the director of Central of Network and Data, and also the director of Key Lab of Network Security and Cryptography, Fujian Normal University, China. He has authored or coauthored over 150 papers in international journals and conferences, including IEEE TRANSACTIONS ON COMPUTER,ACM TRANSACTIONS ON SENSOR NETWORK, IEEE TRANSACTIONS ON RELIABILITY, IEEE TRANSACTIONS ON PARALLEL AND DISTRIBUTED SYSTEMS, INFORMATION SCIENCE, and COMPUTER NETWORK. His current research interests include network and information security, wireless networks and communication, complex networks and systems, and intelligent information in communication networks. Dr. Xu has been invited to act as a PC chair or member at more than 30 international conferences. He is a member of ACM, and a Senior Member of CCF and CIE in China.

Xiaoding Wang received his Ph.D. in College of Mathematics and Informatics from Fujian Normal University in 2016, he is an Associate Professor with the School of Fujian Normal University, China. His main research interests include network optimization and fault tolerance.