ISBN 978-981-14-1684-2
Proceedings of 2019 the 9th International Workshop on Computer Science and Engineering
(WCSE 2019 SUMMER)
Hong Kong, 15-17 June, 2019, pp. 767-775
doi: 10.18178/wcse.2019.06.114

Publishing Correlated Social network Data with Differential Privacy

Siyu Li, Dongran Yu, Xuebo Han, Jie Li, Peng Liu[+], Xianxian Li[+]

Guangxi Key Lab of Multi-source Information Mining & Security, Guangxi Normal University, Guilin, China

liupeng@gxnu.edu.cn, lixx@gxnu.edu.cn

Abstract. The application of social network collects a large amount of user data and sensitive data, which may reveal potential privacy information through analysis. At present, the differential privacy protection model gives a rigorous and quantitative representation and proof of privacy disclosure risk, which greatly guarantees the availability of data. Recently, differential privacy is very popular. However, differential privacy assumes that data sets are independent. In real life, few data sets are completely independent. In social networks, nodes have edges that are related. This paper proposes a solution to use differential privacy on correlation social network data and designs a mechanism for correlation social network data publishing. Reduce the large amount of noise added when the association graph data is published with differential privacy. Considering the degree of correlation between nodes in the graph data, this paper first proposed the degree of association between nodes and calculated the degree of association between each node to calculate the sensitivity of association. The correlation sensitivity is used to determine the noise level in the implementation of differential privacy. Then the hierarchical random graph model is used to add noise satisfying differential privacy to edge connection probability to generate the pending layout. Finally, the feasibility and effectiveness of the method are verified by the statistical analysis of degree distribution, aggregation coefficient and one-dimensional structural entropy.

Keywords: Differential privacy, correlation data, social networks, publishing graph generation model

1. Introduction

Social network has become an indispensable part of our life now, it has greatly changed the pattern of contact between social individuals. Social network applications such as Facebook, LinkedIn, Instagram QQ, WeChat have a large number of customers. People's various behaviors on the Internet produce a lot of social network data. The basic method of social network research is to analyze social network data. The analysis results can be of great commercial and social value to researchers. For example, data managers can conduct sociological research, infectious disease research, business model analysis and so on through relevant data. The value of social network data is gradually reflected.

However, the development of massive data collection and efficient analysis technology of social networks has seriously threatened the privacy security of citizens. In July 2015, a federal judge in the United States ruled that Google had to face privacy litigation because it mixed user data in different Internet products and provided it to advertisers without users' permission. Therefore, how to apply the concept of privacy protection to social network data has become an urgent practical problem. The traditional privacy protection methods based on data anonymization technology have obvious defects. These models cannot provide sufficient security guarantees and to protect against some kinds of background attacks.

This paper proposes a privacy protection method for associating social network graph data based on differential privacy. We consider the sensitivity of the data between the nodes in the connection degree of compute nodes, calculate the sensitivity, the degree of association to use hierarchical random graph model for

[+] Corresponding author. Tel.: 18677334692; 18677334692
E-mail address: liupeng@gxnu.edu.cn; lixx@gxnu.edu.cn

modeling the original data, add Laplace noise in the process of modeling, the established model meet the difference of privacy, finally by establishing the model of production meet the requirements of the privacy issue. This paper focuses on considering the risk of privacy disclosure caused by correlation in social networks, and proposes a protection method that satisfies differential privacy protection by adding related social network data. The concepts of association degree, node sensitivity and association sensitivity are defined to determine the amount of noise to be added, so as to solve the problem of associating rows of data. Hierarchical random graph is used to satisfy differential privacy and improve the structural utility of graph data.

To summarize. We made the following contributions:

- We aim at the privacy protection problem in the generation of social network publishing graph, the definition of the degree of association between nodes and the association sensitivity of the entire network graph is proposed to determine the amount of noise added when differential privacy is used. Solve the problem of associating social networks.
- We use the hierarchical random graph model to represent the structure of a network graph. Then add noise satisfying differential privacy to the connection probability of edges in the hierarchical random graph to express the network graph structure.
- We use the degree distribution, aggregation coefficient and one-dimensional structural entropy statistical analysis method to verify and analyze the structural consistency and effectiveness of the pending layout diagram and the original one through experimental results.

This paper is organized as follows. Section 2 introduces the related work of social network privacy protection. In the third section, the background knowledge of differential privacy, association sensitivity and hierarchical random graph is introduced. In the fourth section, we introduce the algorithm for calculating association sensitivity and the algorithm for adding noise to hierarchical random graph. In section 5, the effectiveness of our algorithm is proved by experiments.

2. Related Works

With the recent widespread recognition of differential privacy, some papers [1] have begun to apply differential privacy to network data from different perspectives. Dwork et al.[2] proposed a method to provide differential privacy protection for general query functions by calculating the real query results and then adding some noise to return the results to users. Hay et al. [3] proposed a post-processing technique to improve the accuracy of query results without sacrificing differential privacy protection. The core idea of the algorithm is the result of the constraint condition of the query sequence after adding noise.Yu et al. [4] proposed the privacy of local difference on the property graph, introduced a new model, and showed how to enhance the privacy protection of the property graph through meaningful privacy. A complete workflow is given in which the input is sensitive graph and the output is real composite graph. Rui Chen et al. [5] adjusted the degree of correlation by introducing an additional parameter and multiplied the privacy protection budget by the number of relevant records to solve the problem of related information in social networks. It also provides an overall solution for non-interactive network data publishing. Qian Xiao et al. [6] proposed a data protection solution to infer the network structure in different private ways. The random graph model statistical hierarchy is used to infer the network structure. Markov chain monte carlo is used to sample HRG structure in model space to guarantee differential privacy. Zhu et al. [7] proposed an efficient solution for related differential privacy and designed a relevant data publishing mechanism, so as to reduce the large amount of noise added to the relevant data set when differential privacy publishing was adopted. Zhan Qin et al. [8] proposed technologies to ensure individual differential privacy while collecting structural information and generating representative composite social graph.

Samarati and Sweeny proposed a new privacy protection method as offered to data publishing k-anonymity [9] in 2002. The k-anonymous model divides the published data into several equivalence classes, and requires that at least k-1 records have the same attributes as any record in the equivalence class, so that the data attacker can only identify the target data with probability less than or equal to 1/ K. Subgraph k-anonymous [10] [11] [12] method: by adding or removing edges or nodes in the social network graph,

attackers can only identify the subgraph structure of the target node with a probability less than 1/k. In 2012, Boldi et al. published a literature on improving the traditional k-anonymous model on VLDB [13] and proposed a method to introduce uncertainty to the local edges of the original social network, so that the entropy of node degree distribution in the published uncertain social network is greater than the log base 2 of the parameter K. In 2017, Wang et al. [14] proposed an effective anonymous method of social network to protect the community structure. In 2010, Hay et al. proposed a differential privacy model for the distribution of social network degree [15] in their paper, which proposed a differential privacy protection method combining hierarchical sum and least square method in order to improve the accuracy of query results without reducing the privacy intensity. In 2013, Wang [16] et al. proposed a differential privacy protection method by perturbing the eigenvalues of the graph and the adjacency matrix corresponding to the eigenvectors. [17] in 2018. In 2016, wang et al. [18] proposed an effective anonymous method of social network to protect community structure. Lei et al. [19] proposed a mechanism based on trust to achieve collaborative privacy management in 2018. Cao et al. [20] proposed in 2017 the potential loss of privacy under the time correlation of traditional differential privacy mechanism under the condition of continuous data release.

Associated in this article, we are considering the difference of social network data privacy protection method, reference HRG [6] model represents a graph structure, the model of HRG, network diagram can be converted to HRG diagram for differential treatment is the structure of the whole figure difference, which in turn is to protect the original data, have retained the original structure characteristics, the purpose of improving the availability of data. The purpose of this paper is to reduce the amount of noise added to related social network data and better protection structure effectiveness. Of association degree between nodes and the network graph associated sensitivity, using random figure said network diagram structure, level in hierarchy random graph to add meet noise difference of privacy, which released after dealing with the difference of privacy of purification figure, at the same time, does not destroy the characteristics of network structure make it can be used for data analysis and mining.

3. Preliminaries

3.1. Differential Privacy

Differential privacy is a strict mathematical framework designed to protect the privacy of sensitive user information in statistical databases. Differential privacy is based on adding random noise to the query results to ensure that the user's privacy risk is not increased by whether or not it is in the database (limited by the privacy budget ϵ). A formal definition of Differential Privacy is as follows:

Definition 1 (ε- Differential Privacy): A mechanism M gives ε-*differential privacy* for any pair of D and D′, and for every set of outcomes Ω, the randomized mechanism M satisfies:

$$\frac{P_r(M(D)\in\Omega)}{P_r(M(D')\in\Omega)}\le \exp(\varepsilon) \qquad (1)$$

The parameter ε, called the privacy budget, represents the degree of privacy offered. Intuitively, a lower value of ε implies stronger privacy guarantee and a larger perturbation noise, and a higher value of ε implies a weaker privacy guarantee while possibly achieving higher accuracy.

Mechanism M is associated with the *global sensitivity*. This measures the maximal change on the result of query Q when removing one record from the dataset D.

Definition 2 (*Global Sensitivity*): For Q: D → R, the *global sensitivity* of Q is defined as:

$$GS=\max_{D,D'}\left\|Q(D)-Q(D')\right\|_1 \qquad (2)$$

To satisfy the differential privacy definition, two mechanisms are usually utilized: the Laplace mechanism and the Exponential mechanism. Between these, the Laplace mechanism is suitable for numeric output and relies on the strategy of adding Laplace noise (denoted as Laplace(·)) to the query result. It is formally defined as:

$$M(D)=Q(D)+Laplace(\frac{GS}{\varepsilon}) \qquad (3)$$

provides the ε-differential privacy.

3.2. Correlation sensitivity

Definition 3.1 (Degree of correlation): An Definition of association degree: suppose that I and j are related to each other, and their correlationship is expressed by association degree , the range is 0~1, and set a threshold a, >a, when the association degree is less than the set threshold, it is not considered, set to 0. Association degree represents the influence of one record on another record. The higher the association degree value is, the stronger the association degree of the record will be. Therefore, the formula for calculating the degree of association between nodes is as follows. The degree of association is calculated by degrees. Two friends share more friends and the degree of association is relatively high.

The correlation degree between nodes is expressed as:

$$\delta_{i,j} = \frac{|(x_1^i,...,x_m^i) \cap (x_1^j,...,x_n^j)| + 1}{2\max(\deg(x^i), \deg(x^j))} \qquad (4)$$

Where, the deg (xi) represents the degree of node xi. (Xin) represents the neighbor node set of node I, and (XJN) represents the neighbor node set of node j.

Definition 3.2 (*Node sensitivity*): Association The correlationship between nodes is known as the association information and the data set in which it resides is defined as the correlation data set. In social network data, a node is associated with its neighbor node to which it is connected.

Record sensitivity definition: for a given and a query Q, the record sensitivity is

$$CS_i = \sum_{j=0}^{n} |\delta ij|(\| Q(D^j)) - Q(D^{-j}\|) \qquad (5)$$

Definition 3.3 (*Correlation sensitivity*): Association sensitivity definition: for a query function Q. Correlation sensitivity is determined by the maximum record.

$$CS_q = \max_{i \in q}(CS_i) \qquad (6)$$

Where q is the record set of all records that respond to query q.

3.3. Hierarchical random graph model

HRG hierarchy represented by binary tree, the binary tree leaf nodes of the network node, HRG given binary tree each internal node r probability of a connection, it says to r for the root of the left subtree Lr corresponding right subtree of the Rr is the probability of an edge, it can reflect about the connection strength between two groups (community). The larger the Pr value is, the closer the connection will be. Pr can be expressed as:

$$p_r = |e_r| / (|n_{L_r}| \cdot |n_{R_r}|) \qquad (7)$$

Where, $|e_r|$ represents the number of connecting edges of left and right subtrees of internal node r, L_r and R_r respectively represents the left and right subtrees of node r, n_{L_r} represents the number of leaf nodes in the left subtree, and n_{R_r} represents the number of middle nodes in the right subtrees.

4. Rcorrelation Social Network Differential Privacy Algorithrm

In order to solve the privacy protection problem of related social network data, this paper proposes the definition of association degree, node sensitivity and association sensitivity. Algorithm 1 is used to calculate the association sensitivity of the association network graph. Algorithm 2 uses hierarchical random graph model to add noise. Algorithm 3 publishes the composite diagram.

4.1. Calculate correlation sensitivity processing

First, the original network graph is input, and the association degree of each node is calculated according to the formula for calculating the association degree between nodes. After calculating the sensitivity of each node, the maximum value is selected as the correlation sensitivity of the network graph by comparing the sensitivity of each node. The specific content is as follows in Table 1:Algorithm1:

Table 1: Algorithm 1

Algorithm 1: calculate the degree of association between two nodes, and then calculate the sensitivity of association
Input: original network graph G
Output: correlation sensitivity CS of network graph
1: for vi in G
2: calculate the degree of association of each node according to formula (4)
3: calculate node sensitivity according to formula (5)
4: select the maximum value from the node sensitivity as the correlation sensitivity

4.2. Add noise to hierarchical random graph

Hierarchical random graph is used to represent a network structure, because a network graph can generate multiple HRGS, so we need the sampling algorithm to select a sample that can best represent the network graph structure from multiple HRGS. We use Metropolis sampling process, first randomly select a tree $T_0 \in T$ to initialize the Markov chain, T is the collection of all possible trees. Then do the loop operation of the second step: randomly select T_{i-1} a neighbor tree T 'and update it in the following way:

$$Ti = \begin{cases} T' & probabilityof\ \alpha \\ T_{i-1} & probabilityof\ \alpha \end{cases} \tag{8}$$

Where, α is the acceptance probability, and the expression is $\alpha = \min\left(1, \frac{\exp(\log L(T'))}{\exp(\log L(T_{i-1}))}\right)$, satisfies $0 < \alpha < 1$, $\log L(T)$ as the standard to measure HRG. $\log L(T, \{p_r\}) = -\sum_{r \in T} n_{L_r} n_{R_r} h(p_r)$ .Select the optimal sample and add Laplace noise to the connection probability of the optimal sample tree. The specific algorithm is as follows in Table 2:Algorithm2:

Table 2 Algorithm 2

Algorithm 2: use hierarchical random graph to add noise
Input: original network graph G, privacy budget, association sensitivity CS, MCMC steady-state threshold
Output: HRG tree diagram with noise added
1: randomly select a tree T to initialize Markov chain；
2: loop through step I in Markov chain：
A) randomly select an internal node r in T_{i-1} ; A construction tree of subtree r is randomly selected as T_{i-1} neighbor tree T;
B) select T 'as T_i the acceptance probability $\min\left(1, \frac{\exp(\log L(T'))}{\exp(\log L(T_{i-1}))}\right)$;
C) if the difference value of logL (T) on Markov chain is < θ, steady state is reached, step 3 is executed; otherwise, step I +1 is executed by jumping back to step 1.
3. Sampling tree is selected in the tree set generated by steady-state Markov chain;
4: return a sample tree S;
5:Calculate the connection probability of internal node r* after adding noise P_{r*} =

$$\min\left\{1, \frac{e_{r^*} + lap(\frac{1}{\varepsilon})}{n_L \cdot n_R}\right\};$$

6: The left subtree of r* is denoted as r_L . Laplace (denoising) (G, S, r_L, ε);

7:The right subtree of r* is denoted as r_R . Laplac(denosing) (G, S, r_R, ε)

8: return the HRG tree diagram after noise addition $\overline{S}$

4.3. Pulish composite diagram

Finally, the purification diagram is generated by placing edges in the network diagram according to the connection probability between noise-adding nodes in Table 3:Algorithm3.

Table 3 Algorithm 3

Algorithm 3: publish composite graph
Input: HRG tree diagram with noise added
Output: purified network diagram
1. HRG tree diagram is represented by matrix;
2: put edges between I and j of the network graph with the connection probability of nodes;
3: output purified network diagram

5. Experiments

In order to measure the performance of our proposed algorithm, we must consider the effectiveness of the proposed method of data privacy protection for social networks using association sensitivity calculation. This section mainly introduces the experimental environment Settings and datasets of the differential privacy protection method for associated social networks.

5.1. Experimental environment and data

In this article, we use the Pycharm2.17.3 development tool. A series of experiments were carried out on a computer with 8.00gb RAM and processor Intel (R) Core (TM) i5-6500 cpu@3.20GHZ. The operating system is Window 7. This paper uses three real datasets to verify the effect of association sensitivity differential privacy method on the utility of datasets.

(1)the ego - Twitter data set (http://snap.stanford.edu/data/index.html) extract 813 nodes and 1768, data from Twitter's social network, data sets to users as a node, use correlationship as a side, the same before the experiment will be treated as data set into a simple undirected graph.

(2)theego - Gplus data set (http://snap.stanford.edu/data/index.html), extract 2076 nodes and 5673, data from Google's social circle.Since the original graph is a directed graph, we now treat the previous network graph as undirected graph before using it.

(3)the theego-Facebook dataset (http://snap.stanford.edu/data/index.html), is a collection of anonymous data from the Facebook, extracted from the data set 2455 nodes and 6823 side.chart.

5.2. Metrics

Degree Distribution:The aggregation coefficient represents the degree to which other nodes in a network are connected to one another.

Definition 5.2.2 (aggregation coefficient) the aggregation coefficient of nodes' is defined as follows:

$$C_i = \frac{2e_i}{d_i(d_i - 1)} \qquad (9)$$

Where, d_i represents the number of neighbors of the node v_i, and e_i represents the number of d_i neighbor correlations existing in each neighbor.

Entropy of one-dimensional structure: Shannon put forward the concept of information entropy in 1948, and structural entropy is a structural information measurement standard based on Shannon's entropy. denoted as . The structural entropy of the graph is given as a graph G, which represents the minimum amount of information known by the coding of nodes in the random walk graph G. The greater the entropy of a graph, the greater the uncertainty of its structure. In this paper, the entropy of one dimensional structure of undirected graph is applied.

$$H^1(G) = H(\frac{d_1}{vol(G)},...,\frac{d_1}{vol(G)}) = -\sum_{i=1}^{n} \frac{d_i}{vol(G)} \log_2 \frac{d_i}{vol(G)} \tag{10}$$

Where vol (G) is called the volume of figure G and $vol(G) = \sum_{v \in G} d_v$

5.3. Experiment result analysis

The This section mainly compares and analyzes the data utility of the three social network data sets in this paper after the privacy protection method with the definition of association sensitivity is added with noise disturbance.Experiment link： https://pan.baidu.com/s/1pA3kH2YSl9lgBYqAuK0Cqw Extraction code： y8fw

1. aggregation coefficient represents the density of nodes in network.

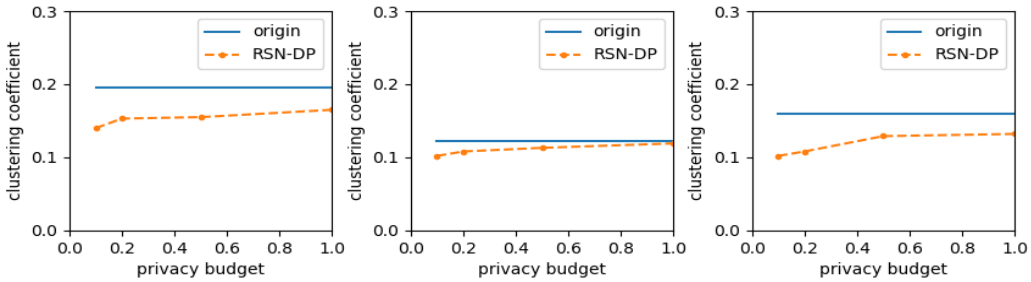


Fig. 1: aggregation coefficient under different privacy budgets

Fig. 1 shows the experimental results of three data sets ego-facebook, ego-gplusand ego-twitter aggregation coefficients under different privacy budgets are respectively shown. The results show that the aggregation coefficient is close to the original graph, and the protection intensity of differential privacy decreases with the increase of privacy budget. In this paper, the better the retention of the average aggregation coefficient of the rsn-dp algorithm is.

2. The structural entropy expresses the uncertsinty of the graph structure,and the higher is the entropy,the graph's structural certainty will be the worse.

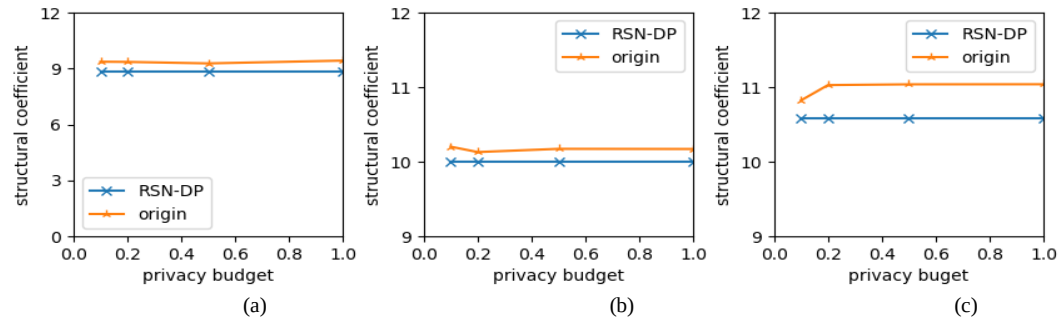
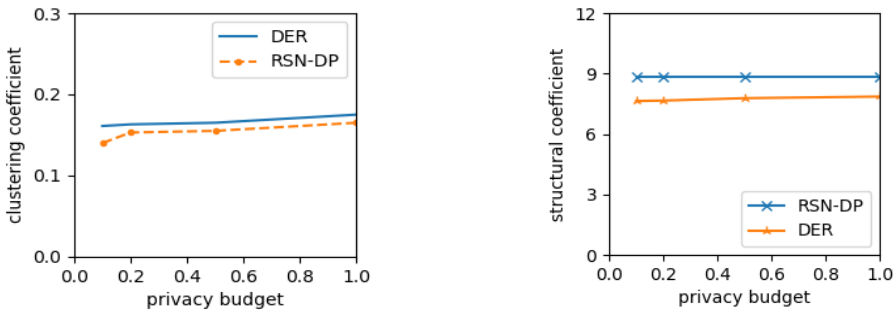


Fig 2: structure entropy under different privacy bud

Fig. 2 shows the experimental results of graph structure performance of the three data sets under different privacy budgets.As the privacy budget increases gradually, the structural uncertainty of the graph becomes stronger.

3. Compared with the method in [5], the experimental results are shown in the figure.



(a) (b)

Fig. 3: compare experiment

6. Conclusion

In this paper, we propose an optimization algorithm. Considering the differential privacy of related social network data, the correlation sensitivity is defined to replace the previous work and the correlation coefficient and the global sensitivity of privacy are multiplied directly, so as to reduce the noise. Then, a non-distributed noise-adding model is designed, which USES hierarchical random graph model to represent the network graph structure. Through experimental analysis, we can draw the conclusion that with the increase of privacy budget, the effectiveness of our algorithm is better. For future work, you can extend this to consider the association of property diagrams.

7. References

[1] Wang J , Liu S , Li Y . A review of differential privacy in individual data release[M]. Taylor & Francis, Inc. 2015.

[2] Proserpio D , Goldberg S , Mcsherry F . Calibrating Data to Sensitivity in Private Data Analysis[J]. 2012.

[3] Hay M , Rastogi V , Miklau G , et al. Boosting the accuracy of differentially private histograms through consistency[J]. Proceedings of the VLDB Endowment, 2010, 3(1-2):1021-1032.

[4] Jorgensen Z , Yu T , Cormode G . Publishing Attributed Social Graphs with Formal Privacy Guarantees[J]. 2016.

[5] Chen, Rui, Fung, et al. Correlated network data publication via differential privacy[J]. Vldb Journal — the International Journal on Very Large Data Bases, 2014, 23(4):653-676.

[6] Xiao Q , Chen R , Tan K L . [ACM Press the 20th ACM SIGKDD international conference - New York, New York, USA (2014.08.24-2014.08.27)] Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining - KDD \"14 - Differentially private network data release via structural inference[J]. 2014:911-920.

[7] Zhu T , Xiong P , Li G , et al. Correlated Differential Privacy: Hiding Information in Non-IID Data Set[J]. IEEE Transactions on Information Forensics & Security, 2014, 10(2):229-242.

[8] Qin Z , Yu T , Yang Y , et al. [ACM Press the 2017 ACM SIGSAC Conference - Dallas, Texas, USA (2017.10.30-2017.11.03)] Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security, - CCS \"17 - Generating Synthetic Decentralized Social Graphs with Local Differential Privacy[J]. 2017:425-438.

[9] Sweeney L . k-ANONYMITY:[J]. International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems, 2008, 10(05):557-570.

[10] Liu K , Terzi E . Towards identity anonymization on graphs[J]. //ACM SIGMOD international conference on Management of Data .ACM, 2008:93-106.

[11] Zhou B , Pei J . Preserving Privacy in Social Networks Against Neighborhood Attacks.[C]// IEEE International Conference on Data Engineering. IEEE, 2008.

[12] Yuan M , Chen L , Yu P S . Personalized privacy protection in social networks[J]. Proceedings of the VLDB Endowment, 2010, 4(2):141-150.

[13] Boldi P , Bonchi F , Gionis A , et al. Injecting Uncertainty in Graphs for Identity Obfuscation[J]. Proceedings of the VLDB Endowment, 2012, 5(11):1376-1387.

[14] Hay M , Rastogi V , Miklau G , et al. Boosting the accuracy of differentially private histograms through consistency[J]. Proceedings of the VLDB Endowment, 2010, 3(1-2):1021-1032.

[15] Wang H , Liu P , Lin S , et al. A Local-Perturbation Anonymizing Approach to Preserving Community Structure in Released Social Networks[J]. 2016.

[16] Wang Y , Wu X , Wu L . Differential privacy preserving spectral graph analysis.[M]// Advances in Knowledge Discovery and Data Mining. 2013.

[17] Wang Q , Zhang Y , Lu X , et al. Real-Time and Spatio-Temporal Crowd-Sourced Social Network Data Publishing with Differential Privacy[J]. IEEE Transactions on Dependable and Secure Computing, 2018, 15(4):591-606.

[18] Wang H , Liu P , Lin S , et al. A Local-Perturbation Anonymizing Approach to Preserving Community Structure in Released Social Networks[J]. 2016.

[19] Lei X , Chunxiao J , Nengqiang H , et al. Trust-based Collaborative Privacy Management in Online Social Networks[J]. IEEE Transactions on Information Forensics and Security, 2018:1-1.

[20] Cao Y , Yoshikawa M , Xiao Y , et al. Quantifying Differential Privacy under Temporal Correlations[C]// IEEE International Conference on Data Engineering. Proc Int Conf Data Eng, 2017.