Differentially Private Social Graph Publishing for Community Detection

Xuebin Ma(✉), Jingyu Yang, and Shengyi Guan

School of Computer Science, Inner Mongolia Key Laboratory of Wireless Networking and Mobile Computing, Inner Mongolia University, Hohhot, China
csmaxuebin@imu.edu.cn

Abstract. Social networks typically include a community structure, and the connections between nodes within the same community are very close; however, the connections between communities are sparse. In this study, we analyze the main challenges behind the problem and then resolve it using differential privacy. First, we choose the Louvain algorithm as a benchmark community detection algorithm for the algorithmic perturbation scheme. We introduce an exponential mechanism that uses modularity as a score. Secondly, by transforming each community into a hierarchical random graph model, and its edge connection probability is noisy by differential privacy mechanism to ensure the security of relevant information in the protected community.

Keywords: Differential privacy · Community detection · Social network

1 Introduction

Techniques for identifying the groupings in social networks and then analyzing these groupings for further use have become a key research topic in sociology—this is referred to as "community detection". Through community detection, we can reduce both the network size and the computational complexity of the algorithm used to process it, thereby improving the accuracy of the analysis. However, most of the methods are performed without privacy protection, and the results of community detection are output in the form of the node-set. In order to protect the privacy of users, it is necessary to protect the privacy of community detection.

Existing social network differential privacy protection schemes focus on a centralized model; they assume that third-party data collectors who possess the information are trustworthy, which is a practical assumption of real-world applications. Therefore, we use a local differential privacy (LDP) model to protect the privacy of social networks, by releasing the sanitized graph at local devices after differential privacy processing. The published data mask the interconnections between nodes and preserve the characteristics of the network structure, enabling researchers to achieve a reasonable balance between the utility of the algorithm and its ability to protect privacy when data are used for feature analysis and data mining.

N. Park et al. (Eds.): SecureComm 2020, LNICST 336, pp. 208–214, 2020.
https://doi.org/10.1007/978-3-030-63095-9_11

2 Related Work

2.1 Local Differential Privacy

The differential privacy protection methods applied to social networks can be roughly divided according to two approaches. The first approach focuses on publishing certain types of noisy mining results; these include degree distributions, subgraph counts, frequent graphics patterns, and cut queries [4, 5]. This approach uses the properties of the original graph for general purposes, perturbs the graph, and publishes the aggregated results. It is theoretically proven that noise addition ensures strong privacy preservation. The second approach is to publish the entire social network for general purposes [6, 7]. These methods differ in the intermediate structures used for publishing and the corresponding definitions of differential privacy.

2.2 Community Detection with Differential Privacy

The task of finding node groups using connection relationships in the network is referred to as community detection. The Louvain algorithm [9] is based on multi-level optimization modularity and performs well in terms of efficiency and effectiveness. Moreover, the Louvain algorithm can identify hierarchical community structures; thus, it is considered one of the best community detection algorithms.

Recently, researchers have applied differential privacy for community detection. Nguyen et al. [10] chose the Louvain method as the backend community detection method of the input perturbation scheme and proposed the LouvainDP method. Ye et al. [11] proposed the first LDP-enabled graph metric estimation framework for a variety of graph analysis tasks, which address data correlation among nodes by two efficient perturbation algorithms based on adjacency bit vector and node degree.

3 Preliminaries

3.1 Louvain Algorithm

The Louvain algorithm [9] is based on multi-level optimization modularity, which is efficient to identify hierarchical community structures. According to [9], when assigning node i to a community, the modularity of the community changes as

$$\begin{aligned}\Delta Q &= \left[\frac{\sum_{in}+k_{i,in}}{2m} - \left(\frac{\sum_{tot}+k_i}{2m}\right)^2\right] - \left[\frac{\sum_{in}}{2m} - \left(\frac{\sum_{tot}}{2m}\right)^2 - \left(\frac{k_i}{2m}\right)^2\right] \\ &= \frac{1}{2m}\left(k_{i,in} - \frac{\sum_{tot} k_i}{m}\right) = \frac{k_{i,in}}{2m} - \frac{\sum_{tot} k_i}{2m^2}. \end{aligned} \tag{1}$$

3.2 Local Differential Privacy

Definition 1 (**ϵ-Differential Privacy**)**:** Given a random algorithm $\mathcal{A}$, let S represent the set of all output spaces of $\mathcal{A}$ on the two neighbor graphs G_1 and G_2(which differ at most one element). The algorithm $\mathcal{A}$ satisfies ϵ-differential privacy if:

$$\Pr[\mathcal{A}(G_1) \in S] \leq e^{\epsilon} \times \Pr[\mathcal{A}(G_2) \in S] \tag{2}$$

Theorem 1 (Laplace Mechanism): For any function $f : \mathrm{G} \rightarrow \mathbb{R}^{\mathrm{d}}$, the mechanism $\mathcal{A}$

$$AG = fG + \left(Lap_1\left(\frac{\Delta f}{\epsilon}\right), \ldots, Lap_d\left(\frac{\Delta f}{\epsilon}\right)\right) \tag{3}$$

provides ϵ-differential privacy, where $Lap_i\left(\frac{\Delta f}{\epsilon}\right)$ represents Laplacian independent and identically distributed variable samples with scale parameter $\frac{\Delta f}{\epsilon}$.

Theorem 2 (Exponential Mechanism): Given a function f : (G × OS) → $\mathbb{R}$, where OS is output space. For a graph G, the mechanism $\mathcal{A}$ that samples an output O with a probability proportional to $\exp\left(\frac{\epsilon \cdot \mathrm{f(G,O)}}{2\Delta \mathrm{f}}\right)$ satisfies ϵ-differential privacy.

Theorem 3 (Sequential Composition): Let each $\mathcal{A}_{\mathrm{i}}$ provide ϵ_{i}-differential privacy. A sequence $\mathcal{A}_{\mathrm{i}}(\mathrm{G})$ over the entire graph G provides $\sum \epsilon_{\mathrm{i}}$-differential privacy.

4 Problem Solution

4.1 Differentially Private Louvain Algorithm

We propose a solution to the privacy problems in community detection. Our scheme is divided into two phases. First, the social network is partitioned into multiple independent communities by adopting the community detection algorithm. Then, the privacy of the edges within each independent community is protected.

Algorithm 1: Differentially Private Louvain Algorithm

Input: Input graph G, privacy parameter ϵ_1
Output: a private partition set $\{P_1,...,P_k\}$

1 Calculate the sensitivity Δf
2 Randomly select initial node sequences and each node as a partition
3 **for** each node i in sequences S **do**
4 **for** neighbor partition P_k from partition set **do**
5 compute the modular gain ΔQ_k
6 $\Delta Q_{max} = \Delta Q_k$ with the probability min $\{1, \frac{\exp\left(\frac{\epsilon_1}{2\Delta f}\Delta Q_k\right)}{\exp\left(\frac{\epsilon_1}{2\Delta f}\Delta Q_{max}\right)}\}$
7 Record the partition P_i in which ΔQ_{max} is obtained
8 **end for**
9 move node i into the partition P_i
10 if the partition of all nodes no longer changes
11 **return** private partition set $\{P_1,...,P_k\}$
12 **end for**

Algorithm 1 first calculates the sensitivity of the social graph according to its number of nodes (Line 1), and we will explain in detail how to calculate Δf later. Based on

the first phase of the Louvain algorithm, each node in the original graph is treated as a separate community (Line 2). Then, according to the node sequence, mining the neighboring nodes/communities of each node (lines 3–4). Then moving the node to different communities and calculating the current modularity gain. Finally, we introduce the exponential mechanism, and the maximum modularity gain is selected, otherwise, it is unchanged (lines 5–9). When the movement of all nodes no longer causes changes of the modularity gain, the first round is completed and the results of the first round of community detection are returned (lines 10–12).

4.2 Edge Probability Perturbation

Community detection makes edge connection within the same community more salient and therefore requires additional protection. We use the same model in [15], which converts each community into an HRG model, then combined the generated HRG model with the edge-connection probability by adding Laplace noise in Algorithm 2.

Algorithm 2: Edge Probability Perturbation

Input: Input partition set $\{P_1,...,P_k\}$, privacy parameter ϵ_2
Output: Sanitized graph $\widetilde{G}$
1 **for** each partition P_i in set $\{P_1,...,P_k\}$ **do**
2 Convert to HRG model T_i
3 **for** each internal node r of T_i **do**
4 Calculate noisy probability $\widetilde{p_r} = \min\left\{1, \frac{e_r + Lap\left(\frac{1}{\epsilon_2}\right)}{n_{L_r} * n_{R_r}}\right\}$
5 **end for**
6 **for** any two nodes i, j of P_i **do**
7 Find the lowest common ancestor r of i and j
8 Place an edge in P_i between i and j with independent probability $\widetilde{p_r}$
9 **end for**
10 **end for**
11 Connect partition $\widetilde{P_1},...,\widetilde{P_k}$
12 **return** Sanitized graph $\widetilde{G}$

After we convert the community into an HRG model, we calculate the connection probability of each internal node separately. Further, we introduce the Laplacian mechanism. Subsequently, for any two nodes i and j in the community, we find the lowest common ancestor r and establish a connection between the two nodes i and j using the connection probability of the internal node r. Because the inter-community edges are relatively sparse and have low correlations, these direct connections do not provide additional privacy protection.

4.3 Sensitivity Analysis

The sensitivity Δf should be analyzed to complete the selection probability equation, where G' is the neighbor of G. The neighbor of a graph is the graph obtained by changing only one edge. Because the addition or removal operations are similar, only the former is considered in our proof. There are two cases to consider: (1) The connection is an edge inside the community P. (2) The connection is an edge between the community P and S. Finally, we obtain $\Delta f \leq \frac{3}{m}$.

5 Experiment Evaluation

5.1 Experimental Setup

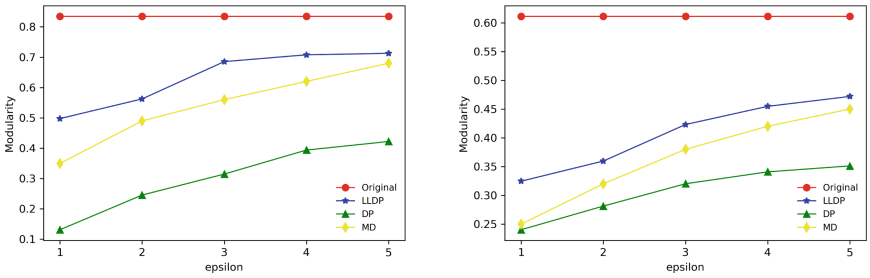
For comparison purposes, two techniques that are similar to our method were implemented as references. They are the basic differential privacy algorithms for the HRG model, which use the same privacy criteria as [13] and the algorithm perturbation presented in [10]; the results of the previously centralized differential privacy scheme and the algorithm perturbation scheme and our proposed scheme are labeled as "DP", "MD" and "LLDP" respectively. We performed experiments on two real datasets to evaluate our algorithm. The details of datasets are shown in Table 1.

Table 1. Statistics of the datasets.

Datasets	Nodes	Edges	Average clustering coefficient
Ego-Facebook [12]	4039	88234	0.6055
Enron [13]	36692	183831	0.4970

5.2 Experiment Evaluation of Community Detection

The real social network dataset we chose did not have standard community detection results. Thus, we chose the output of the Louvain algorithm as a standard control because the evaluation of the data had been performed in [15], and the Louvain method had been proven there to provide high-quality results.



(a) Ego-Facebook (b) Enron

Fig. 1. The modularity of two social network datasets under different ϵ

The partitioning results of different privacy budgets are shown in Fig. 1. For both datasets, the results given by our algorithm increased with the increase of the privacy budget and gradually stabilized after reaching a certain value. This provides an effective reference for the selection of a privacy budget. In general, the effectiveness of algorithm

perturbations is higher than that of input perturbations. The input disturbance used for comparison also followed a similar trend; however, the overall value was low, which is significantly different from its typical value. Newman [8] suggested that the value of Q in a general network is between 0.3 and 0.7, which can explain a good community structure. Therefore, although the modularity of the results obtained by our algorithm was lower than the real situation, it still retained an effective community structure.

6 Conclusion

We analyzed the privacy problems of community detection can lead to and proposed a differentially private detection procedure based on the Louvain algorithm. Moreover, we proposed to further protect the relational data within the community by converting the individual community into a subgraph of an HRG model and subsequently calculating the edge connection probability by adding Laplacian noise. Experimental results indicated an improved performance on real data.

Acknowledgments. This paper is supported by Inner Mongolia Natural Science Foundation (Grant No. 2018MS06026) and the Science and Technology Program of Inner Mongolia Autonomous Region (Grant No. 2019GG116).

References

1. Newman, M.: Social networks. Oxford Scholarship Online (2018)
2. Kasiviswanathan, S.P., Lee, H.K., Nissim, K., Raskhodnikova, S., Smith, A.: What can we learn privately? SIAM J. Comput. **40**(3), 793–826 (2011)
3. Dwork, C.: Differential privacy: a survey of results. In: Agrawal, M., Du, D., Duan, Z., Li, A. (eds.) Theory and Applications of Models of Computation. Lecture Notes in Computer Science, vol. 4978, pp. 1–19. Springer, Berlin, Heidelberg (2008). https://doi.org/10.1007/978-3-540-79228-4_1
4. Hardt, M., Roth, A.: Beating randomized response on incoherent matrices. In: Proceedings of the 44th symposium on Theory of Computing - STOC'12 (2012)
5. Cai, Z., He, Z., Guan, X., Li, Y.: Collective data-sanitization for preventing sensitive information inference attacks in social networks. IEEE Trans. Dependable Secure Comput. **15**, 1 (2016)
6. Proserpio, D., Goldberg, S., McSherry, F.: A workflow for differentially-private graph synthesis. In: Proceedings of the 2012 ACM Workshop on Workshop on Online Social Networks - WOSN'12 (2012)
7. Wang, Y., Wu, X., Wu, L.: Differential privacy preserving spectral graph analysis. In: Pei, J., Tseng, V.S., Cao, L., Motoda, H., Xu, G. (eds.) Advances in Knowledge Discovery and Data Mining. Lecture Notes in Computer Science, vol. 7819, pp. 329–340. Springer, Berlin, Heidelberg (2013). https://doi.org/10.1007/978-3-642-37456-2_28
8. Newman, M.E.J., Girvan, M.: Finding and evaluating community structure in networks. Phys. Rev. E **69**(2), 026113 (2004)
9. Blondel, V.D., Guillaume, J.-L., Lambiotte, R., Lefebvre, E.: Fast unfolding of communities in large networks. J. Stat. Mech: Theory Exp. **2008**(10), P10008 (2008)

10. Nguyen, H.H., Imine, A., Rusinowitch, M.: Detecting communities under differential privacy. In: Proceedings of the 2016 ACM on Workshop on Privacy in the Electronic Society - WPES'16 (2016)
11. Ye, Q., Hu, H., Au, M.H., Meng, X., Xiao, X.: Towards locally differentially private generic graph metric estimation. In: 2020 IEEE 36th International Conference on Data Engineering (ICDE), Dallas, TX, USA, pp. 1922–1925 (2020)
12. McAuley, J., Leskovec, J.: Learning to discover social circles in ego networks. In: NIPS 2012, pp. 539–547
13. Leskovec, J., Lang, K.J., Dasgupta, A., Mahoney, M.W.: Community structure in large networks: natural cluster sizes and the absence of large well-defined clusters. Internet Math. **6**(1), 29–123 (2009)
14. Xiao, Q., Chen, R., Tan, K.-L.: Differentially private network data release via structural inference. In: Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 911–920, ACM (2014)
15. Prat-Pérez, A., Dominguez-Sal, D., Larriba-Pey, J.-L.: High quality, scalable and parallel community detection for large real graphs. In: Proceedings of the 23rd international conference on World Wide Web - WWW'14 (2014)