2-6-2019

De-anonymyzing scale-free social networks by using spectrum partitioning method

Qi Sun
Qufu Normal University

Jiguo Yu
Qufu Normal University

Honglu Jiang
Qufu Normal University

Yixian Chen
Michigan Technological University

Xiuzhen Cheng
The George Washington University

Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 147 (2019) 441–445

Procedia Computer Science

www.elsevier.com/locate/procedia

2018 International Conference on Identification, Information and Knowledge in the Internet of Things, IIKI 2018

De-anonymizing Scale-Free Social Networks by Using Spectrum Partitioning Method

Qi Sun[a], Jiguo Yu[a,*], Honglu Jiang[a], Yixian Chen[b], Xiuzhen Cheng[c]

[a]*School of Information Science and Engineering, Qufu Normal University, Rizhao 276826, China*
[b]*Department of Computer Science, Michigan Technological University, Houghton MI 49931, USA*
[c]*Department of Computer Science, The George Washington University, Washington DC 20052, USA*

Abstract

Social network data is widely shared, forwarded and published to third parties, which led to the risks of privacy disclosure. Even thought the network provider always perturbs the data before publishing it, attackers can still recover anonymous data according to the collected auxiliary information. In this paper, we transform the problem of de-anonymization into node matching problem in graph, and the de-anonymization method can reduce the number of nodes to be matched at each time. In addition, we use spectrum partitioning method to divide the social graph into disjoint subgraphs, and it can effectively be applied to large-scale social networks and executed in parallel by using multiple processors. Through the analysis of the influence of power-law distribution on de-anonymization, we synthetically consider the structural and personal information of users which made the feature information of the user more practical.

Keywords: social networks, de-anonymization, privacy-preserving, graph partition;

1. Introduction

Many online social networks like Facebook, Twitter or Weibo have provided great convenience for people to contact. However, most users do not have the awareness of privacy protection , and their privacy will be obtained by malicious third parties. Therefore, social network data analysis and privacy issues are attracting increasing attention from researchers [1],[2],[3],[4],[5]. The privacy issue is one of the problems to be solved in online social networks. Simply removing the ID can't protect user's privacy.

* Corresponding author. Tel.: +86-633-3980462; fax: +86-633-3980462.
E-mail address: jiguoyu@sina.com

1877-0509
10.1016/j.procs.2019.01.262

To protect the privacy of users in online social networks, researchers have proposed various anonymous protection technologies, including k-anonymity [6] and its variants l-diversity [7] and t-closeness [8], as well as differential privacy which has good performance [9]. Some papers have proved that using auxiliary graphs to de-anonymize anonymous network can obtain user's identity information and attribute information to a certain extent [2],[3],[5],[10]. However, most of the de-anonymization algorithms just consider the structural features of the network and ignore the user attributes, and it can't characterize the node's feature information well.

We are committed to constructing a comprehensive and realistic model to execute the de-anonymization process. Our contribution mainly includes the following three aspects:

- we give the de-anonymization method which considers reducing the number of candidate nodes and use the structural features and attribute values to match two nodes.
- we extract the node degree centrality, weighted node degree centrality and cluster coefficient to describe the feature information of the nodes.
- We use spectral partitioning method to divide large-scale social networks into disjoint sub-networks, which allows our algorithm to execute in parallel on multiple processors.

Outline. The structure of the paper is deployed as follows: In section 2 we introduce the related work. The network model and attack model are given in section 3. Section 4 describes the de-anonymization process in detail. In section 5 we give the simulation evaluations. We summarize the paper in section 6.

2. Related Work

2.1. Anonymizaiton Methods

A simple approach of protecting user's identity is to use anonymous technologies such as removing a user's identification [2],[3],[11]. Although removing a user's ID can hide the true identity of the user, this method can't resist the structure-based de-anonymization attack. In addition, in order to protect the user's identity well, the *k*-anonymous method [6] was proposed. The researchers extended *k*-anonymity and *l*-diversity technologies to further enhance the performance of anonymity.

Recently, researchers used the method of differential privacy to protect relevant data. In [12], Sale et al. proposed a differential privacy graph model to protect the privacy of user relationships (e.g., friends, partners, etc). In [13], Xiao et al. proposed a data sanitization scheme based on differential privacy estimating the probabilities that user contact. In [14],[15], Ji et al. studied the utility and security of the existing graph anonymity techniques. But through a large number of analysis and experiments, the results show that the existing anonymous technologies still can't resist the current graph de-anonymization attacks.

2.2. De-anonymization Attack

Most previous de-anonymization techniques only consider the structural characteristics of the network using the local and global structural features of the network to achieve the de-anonymization. In [2], Narayanan and Shmatikov proposed a scalable and robust de-anonymization method. The attack consists two phases which can re-identify users with low error. In [3], structure-based de-anonymization techniques are applied to the de-anonymization of mobility trajectories. In [10], Ji et al. presented a unified method based on structured de-anonymization technology, which can recover the social network and the mobility trajectory. In [16], Nilizadeh et al. proposed the community-based algorithm which greatly improves the seed-based de-anonymization attacks.

None of the above de-anonymization methods take into account the user's attribute information. In [17], Qian et al. used knowledge graph to model the background of social network graph of attackers. By using the correlation between user attributes, they obtained the users identity and attribute information. In [18], Jiang et al. used SA framework as the network model, and considered the attribute differences of users which can improve the accuracy of matching. In calculating the node similarity, the user's attribute relevance has become an important element.

3. Social Network Model

3.1. Social Network Graph Model

We will model the data in social networks as graph $G = (V, E, W, A)$, where $V = \{i | i \text{ is a node}\}$ represents user set in the social network, $E = \{e_{ij} | i, j \in V\}$ represents the relationship between users, $W = \{w_{ij} | i, j \in V\}$ is the weight set on edges, $A = (v_1, v_2, ...)$ is user's attribute. Specifically, the set of attributes for user i is denoted by A_i. Assume that the anonymous graph released by the data owner is $G_a = (V_a, E_a, W_a, A_a)$. Further we define $n = |V|$ and $m = |E|$ as the number of users in the graph and the number of edges respectively, and $N(i)$ is the neighbor set of the user.

3.2. Attack Model

Online social network providers publish anonymous data G_a to third parties. We assume that the provider is honest and will not reveal additional data to others. Attackers can obtain some information about the user through various means which constructed an auxiliary graph. And the auxiliary graph is represented by $G_u = (V_u, E_u, W_u, A_u)$. Attackers have the ability to access the data of anonymous graph, so that the collected data can be analyzed and utilized. The purpose of the attacker is to use the collected auxiliary information to obtain the user identity in the anonymous graph.

4. Scheme Detail

The main steps of our scheme are as follows. Firstly, we divide a large-scale social graph into smaller subgraphs by using the method in [19]. Secondly, we match the subgraphs of the anonymous and auxiliary graphs. Thirdly, we present a model of matching nodes in the matched subgraphs. Finally, we consider the network structure and attribute information to measure the similarity between nodes.

4.1. Structural Similarity of Nodes

To measure the structural similarity, we extract the degree centrality, weighted degree centrality and clustering coefficient of nodes.

4.1.1. Nodes Degree Centrality

The degree centrality of a node is defined as the number of edges connected to this node. For example, considering the arbitrary social graph $G = (V, E, W, A)$, the degree of node $v \in V$ is $d_v = |N(v)|$, where $|N(v)|$ is the number of v' neighbors.

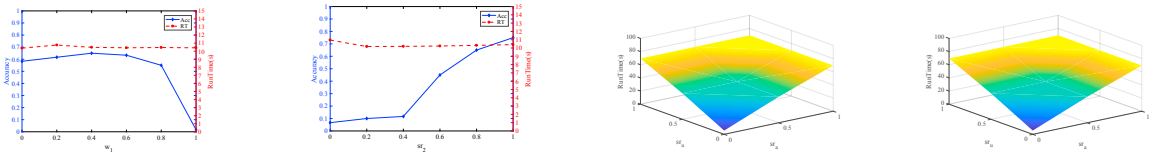
In calculating the degree of centrality of the weighted graph, we use the definition of weighted centrality given in [20]. For node $v \in V$, the degree of centrality of v is defined as:$wd_v = d_v(\frac{\sum_{u \in N(v)} w_{vu}}{d_v})^{\alpha}$,Where α is a positive tuning parameter. When $0 \leqslant \alpha \leqslant 1$, the large degree is regarded to be important; when $\alpha \geqslant 1$, the weight is regarded to be important.

4.1.2. Nodes Clustering Coefficient

The clustering coefficient [21] measures the degree to which nodes tend to cluster together in social graph G. The clustering coefficient C_i of node $v_i \in V$ is defined as the number of edges connected to this node divided by the number of all possible edges between its neighboring nodes. For a directed graph, the number of possible edges between all neighbors $N(v_i)$ of v_i is $k_i(k_{i-1})$, where k_i is the number of neighbors of node v_i. Therefore, the clustering coefficient for directed graph is defined as:

$$C_i = \frac{|e_{ij} : v_i, v_j \in N_i, e_{ij} \in E|}{k_i(k_i - 1)} \tag{1}$$

Through the above introduction, now we extract the structural features of networks. Firstly, for the nodes $v_i \in V_a$ and $v_j \in V_u$, we define their structural feature vectors $S_a(v_i)$, $S_u(v_j)$ as $S_a(v_i) = [d_i, wd_i, C_i]$, $S_u(v_j) = [d_j, wd_j, C_j]$



(a) (b) (c) (d)

Fig. 1. The impact of w_1,sr_2,sr_a and sr_u on de-anonymization accuracy and run time.

respectively. Next, we define the structural similarity S_S between two nodes $v_i \in V_a$ and $v_j \in V_u$ as the cosine similarity between these two vectors:

$$S_S(v_i, v_j) = \frac{S_a(v_i) \cdot S_u(v_j)}{||S_a(v_i)||||S_u(v_j)||} \tag{2}$$

4.2. Attribute Similarity of Nodes

We have given the basic model of social network $G = (V, E, W, A)$, where A is the user's attribute. For each user v_i in the graph, $A_i = (v_{i1}, v_{i2}, ..., v_{im})$ represents its attribute vector, where $v_{ik} = 1$ if user has this attribute; otherwise $v_{ak} = 0$. When calculating the attribute similarity S_A between nodes, we consider the number of common attributes the two nodes have. Specifically, for two nodes $v_i \in V_a$, $v_j \in V_u$, the attribute similarity $S_A(v_i, v_j) = |A_i \bigcap A_j|/|A_i \bigcup A_j|$.

4.3. The Similarity Between Nodes

Now, we define the similarity $Sim(v_i, v_j)$ for $v_i \in V_a$ and $v_j \in V_u$ as follows:

$$Sim(v_i, v_j) = w_1 S_S(v_i, v_j) + w_2 S_A(v_i, v_j) \tag{3}$$

where w_1,w_2 are weight coefficients, and $w_1 + w_2 = 1$. If $w_1 > w_2$, it means that the structural information is more important; vice versa.

4.4. De-anonymization

We consider a practical social network characteristic here, that is, scale-free degree distribution. The node degree in the network follows power-law distributions. Therefore, we select the node with the largest degree of nodes in each subgraph to start our method. The execution of the de-anonymization process is similar to percolation-based method. Firstly, the node with the largest degree in the graph is de-anonymized and then its neighbors. In this process, it just compares the nodes in the neighbors of matched nodes. It can reduce the complexity of the algorithm efficiently.

5. Experiment Evaluations

We conducted experiments on two large social network datasets, Facebook and Google+, in the real world. They are from the Standford Network Analysis Project (SNAP). In the experiments, the default value of parameters w_1,w_2 in the algorithm are $w_1 = w_2 = 0.5$. At the same time, the default settings of sampling frequency for generating anonymized graph and auxiliary graph are $sr_a = sr_u = sr_1 = 0.8$. Assume that the attribute information of nodes in the two networks is obtained by sampling from the original attribute information, and the sampling frequency sr_2 defaults to 0.9. The experimental operating environment is: Intel CoreTM 2.4GHz 8-core CPU 16G ROM.

In Fig. 1(ab), we show the effects of parameters w_1 and sr_2 on the accuracy and runtime. In Fig. 1(a), it verifies the idea that attribute information has a positive effect on de-anonymization algorithm. In Fig. 1(b), we can see that when sr_2 is small, the accuracy of the algorithm is low, because the perturbation of the attribute information is large at this time. With the increase of sr_2, the accuracy of the algorithm also gradually increases, which further illustrates the ideal that the attribute has an auxiliary effect on the accuracy of the algorithm.

In Fig. 1(*cd*), we show the influence of the edge sampling frequencies sr_a and sr_u on the accuracy and running time of the matching. From Fig. 1(*c*) we can see that when sr_a and sr_u are close to 1, the de-anonymization algorithm has a very high accuracy. In Fig. 1(*d*) we calculate the running time of the algorithm at different sampling frequencies.

The experiments on Google+ have similar results. For Facebook, the default value of weight is 1 when calculating the weight of the edge. When calculating the edge weight of Google+, if the edge between two nodes is bidirectional, it is set to 2, otherwise it is set to 1.

6. Conclusion

In this paper, we construct a comprehensive and realistic social network graph model, in which not only the structural features of graphs but also the user attribute information are considered. In addition, we use the graph partitioning method to divide social graph which reduces the scale of the problem effectively and enables the method executed in parallel. We also verified the method through the realistic social networks.

Acknowledgment

This work is supported by NSF of China under Grants 61373027 and 61672321.

References

[1] L. Backstrom, C. Dwork, and J. Kleinberg. (2007) "Wherefore art thou r3579x?:anonymized social networks, hidden patterns, and structural steganography." *International Conference on World Wide Web*: 181–190.
[2] A. Narayanan and V. Shmatikov. (2009) "De-anonymizing social networks." *Security and Privacy, 2009 IEEE Symposium on*: 173–187.
[3] M. Srivatsa and M. Hicks. (2012) "Deanonymizing mobility traces:using social network as a side-channel." *ACM Conference on Computer and Communications Security*: 628–637.
[4] S. Ji, W. Li, M. Srivatsa, and R. Beyah. (2014) "Structural data de-anonymization:quantification, practice, and implications." **7** (4): 1040–1053.
[5] J. Qian, X.-Y. Li, Y. Wang, S. Tang, T. Jung, and Y. Fan. (2017) "Social network de-anonymization: More adversarial knowledge, more users re-identified?" *arXiv preprint arXiv:1710.10998*.
[6] L. Sweeney. (2002) "k-anonymity: A model for protecting privacy." *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems* **10** (5): 557–570.
[7] A. Machanavajjhala, J. Gehrke, D. Kifer, and M. Venkitasubramaniam. (2006) "L-diversity: privacy beyond k-anonymity." *International Conference on Data Engineering*: 24–24.
[8] S. Chester. (2013) "Complexity of social network anonymization." *Social Network Analysis & Mining* **3** (2): 151–166.
[9] C. Dwork. (2006) "Differential privacy." *International Colloquium on Automata, Languages, and Programming*: 1–12.
[10] S. Ji, W. Li, M. Srivatsa, J. S. He, and R. Beyah. (2014) "Structure Based Data De-Anonymization of Social Networks and Mobility Traces." *International Conference on Information Security*: 237-254.
[11] M. Hay, G. Miklau, D. Jensen, D. Towsley, and C. Li. (2010) "Resisting structural re-identification in anonymized social networks." *VLDB Journal* **19** (6): 797–823.
[12] A. Sala, X. Zhao, C. Wilson, H. Zheng, and B. Y. Zhao. (2011) "Sharing graphs using differentially private graph models." *In Proceedings of ACM SIGCOMM Conference on Internet Measurement Conference*: 81–98.
[13] Q. Xiao, R. Chen, and K. L. Tan. (2014) "Differentially private network data release via structural inference." *In Proceedings of ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*: 911–920.
[14] S. Ji, P. Mittal, and R. Beyah. (2017) "Graph data anonymization, de-anonymization attacks, and de-anonymizability quantification: A survey." *IEEE Communications Surveys & Tutorials***19** (2): 1305–1326.
[15] S. Ji, W. Li, P. Mittal, X. Hu, and R. Beyah. (2015) "Secgraph: a uniform and open-source evaluation system for graph data anonymization and de-anonymization." *In Proceedings of Usenix Conference on Security Symposium*: 303–318.
[16] S. Nilizadeh, A. Kapadia, and Y. Y. Ahn. (2014) "Community-enhanced de-anonymization of online social networks." *In Proceedings of the VLDB Endowment*: 537–548.
[17] J. Qian, X. Y. Li, C. Zhang, and L. Chen. (2016) "De-anonymizing social networks and inferring private attributes using knowledge graphs." *In Proceedings of IEEE INFOCOM 2016 - the IEEE International Conference on Computer Communications*: 1–9.
[18] H. Jiang, J. Yu, C. Hu, C. Zhang, and X. Cheng. (2018) "SA framework based de-anonymization of social networks." *Procedia Computer Science* **129**: 358–363.
[19] Ma J , Qiao Y , Hu G , et al. (2017) "De-anonymizing Social Networks with Random Forest Classifier." *IEEE Access*(99):1-1.
[20] T. Opsahl, F. Agneessens, and J. Skvoretz. (2010) "Node centrality in weighted networks: Generalizing degree and shortest paths." *Social Networks* **32** (3): 245–251.
[21] W. DJ and S. SH. (1998) "Collectivedynamics of small-world networks." *Nature*: 440–442.