Research on Social Network Dada Utility of Privacy Anonymity Algorithms

Yuqin Xie [a], Mingchun Zheng [b]

School of Management Science & Engineering, Shandong Normal University, Jinan 250014, China

[a]sdnuxyq@163.com, [b]zhmc163@163.com

Abstract

For preserving privacy of social network users when publishing data, many anonymity methods based on k-anonymity have been proposed. While existing many anonymity algorithms may result in nontrivial utility loss with nodes and edges adding or deleting operations.To address this problem, we present a novel anonymity algorithm with nodes division targeting less data utility loss. The evaluation results show that our method can effectively reduce the data utility loss as compared to other network graph modification anonymous algorithms.

Keywords

Social Network, Data Utility, Node Division, Privacy, Anonymity.

1. Introduction

Nowadays, partly driven by many Web 2.0 applications, more and more social network data have been made publicly available and analyzed in one way or another.[1] The social network data has significant application value for commercial and research purposes. For example, Fig.1 describes the relationship between part of the music audience, by data analysis on the network subgraph ,we can get the common interest groups within the network, eventually excavate potential users information.

However, the social network data often have much privacy information of individuals. [2] So it has become a major concern to trade-offs between the individual's privacy security and the data utility while publishing the social network data.[3]

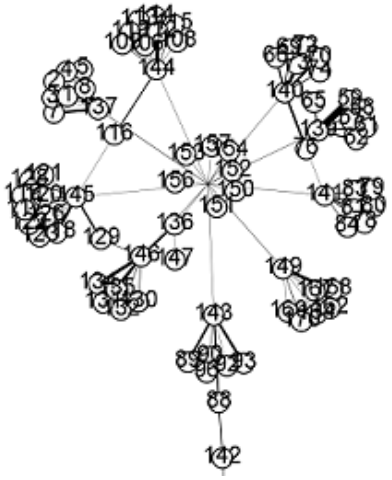

Fig.1 A Last.fm subgraph of music audience

The social networks are modeled as graphs in which nodes and edges correspond to social entities and social links between them respectively, meanwhile a node's label is composed of the corresponding user's attributes, Which are denoted either as sensitive or as non-sensitive.[4] Although various anonymous models have been proposed to preserve the privacy of social network users in existing research[6–8] , the balance between privacy and utility is still a new research topic in the field of social network publishing. What's more, many existing algorithmst may result in nontrivial utility loss with excessive nodes or edges modification operations.

To tackle this issue, we proposed a novel algorithm based on nodes division which shifts anonymous objects from the attribute to some sensitive attribute value, meanwhile explores the different enfluences of nodes on social network topological structure.

2. Model Building

2.1 Basic Knowledge

Definition 2.1 Social network. A social network graph is a four tuple G = (V, E, A, S) as is showed in Fig.2, Where V is a set of nodes of the graph, $E \subseteq V \times V$ is a set of edges between nodes, A is a set of attribute nodes representing all possible values of the privacy attribute, each of which corresponds to a node in a social network. For example, attribute for health-state, pneumonia and influenza as two different property values, form two different attribute nodes, and $S(A_i)$: $A_i \rightarrow S_i$ is an attribute value sensitivity function which is explained in formula (1).

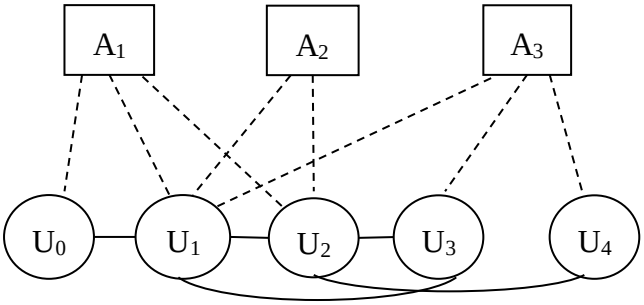

Fig.2 A sample social network graph

Definition 2.2 Structural degree Attack.[5] Given a social network G, its published graph G*, a target entity $t \in V$ and the attacker background knowledge degree of node t, the attacker performs the degree attack by searching for all the vertices in G* that could be mapped to t, i.e., V' = {v∈V*|degree (v) = degree (t)}. If |V'| << |V*|, then t has a high probability to be re-identified.

Definition 2.3 k-degree anonimity.[6] Given a graph G = (V, E) with V = {v_1, v_2, …,v_n} and $d(v_i)$ =|{u ∈V : (u, v_i)∈E}|, and a type of attacker's background knowledge F, the degree sequence of G is defined to be the sequence P=($d(v_1)$, $d(v_2)$, . . . , $d(v_n)$), P can be divided into a group of subsequences [[P[1], ..., P[i_1]], [P[i_1 + 1], ..., P[i_2]], ..., [P[i_m + 1], ..., P[j]]] such that G satisfies k-degree anonymous if for every vertex $v_i \in V$, there exist at least k-1 other vertices in G with the same degree as v_i. In other words, for any subsequences P_y=[P[i_y+1], …,P[i_{y+1}]], P_y satisfies two constraints: (1) All the elements in P_y share the same degree (P[i_y+1].d=P[i_y +2].d=...= P[i_{y+1}].d); (2)P_y has size at least k(i_{y+1} − $i_y \geq$ k).

Definition 2.4 Data utility.[7] The data utility is the availability of data information, which determines the the actual analysis application value of the published social network data. The availability usually contains two aspects: social network users information, social network structure.

Therefore, in order to reduce the loss of data of anonymous data, we comprehensively consider the effect of anonymous operation of these two aspects, design a new anonymous algorithm

2.2 Social network users information analysis

According to the realistic social network, almost every attribute has a rich diversity of values while only some of which are sensitive information. [8]For example, attribute disease status, the privacy degree of HIV is much larger than Cold. But existing methods often specify all values of privacy attribute are uniformly privacy degree and then implement the same anonymization degree. [9]Therefor, these methods have the problem of excessive anonymous for preserving sensitive attribute.

To reduce information less, we make differential protection for diverse sensitive attribute values.

According to each sensitive attribute value has difference sensitivity degree, we put privacy attribute values into three anonymous equivalence group which are denoted by H(high), M(middle), and

L(low) respectively by sensitivity function proposed in fomular (1). When $S(A_i)$=M or $S(A_i)$=H, we consider the attribute value A_i as privacy, and put user node Vi into the protection demond list SV.

$$Sensitivity(x) = \begin{cases} 0, x \le a \\ \dfrac{x-a}{b-a}, a < x \le b \\ 1, x > b \end{cases} \tag{1}$$

$$s.t. \begin{cases} x \le a : S = L(low); \\ a < x \le b : S = M(middle); \\ x > b : S = H(high) \end{cases}$$

Where the critical points a, b is the threshold, for the convenience of the experiment, this article sets threshold based on experience: a=0.6, b=0.8. But for the practical application, it should be performed by statistics and analysis.

2.3 Social network structure analysis

As pointed out in previous sections, the different anonymity oprations have different disruptions on social network structure. The edge or nodes additions(or delections) modifications operations is not effective as them may bring excessive interference to user attributes distribution and social network structure.

In this paper, to measure the structure changes of anonymous graph G* ,we utilize two topological structure indicators APL[4] and CC[4,10], which are essential for graph structure.

$$L = \frac{2}{N(N-1)} \sum_{i \ge j} d_{ij} \tag{2}$$

Where N is the number of network nodes, d_{ij} refers to the distance between node i and j.

$$CC_i = \frac{2E_i}{k_i(k_i - 1)} \tag{3}$$

Where k_i is the neighbor number of node i, E_i is the number of edges between neighbor nodes. If node i have only one neighbor nodes or not (i.e. ki = 1 or ki = 0), the CC_i = 0, obviously $0 \le C_i \le 1$.

2.4 Nodes division

To improve publishing data utility, we designed a novel privacy protection anonymity model in this chapter. The most remarkable of these model ideas is that we use the nodes division instead of nodes or edges addition/deletion etc. graph modification operations.

The definition and the division principle of nodes division which we put forward is reference the theory of biological cell division combined with the relevant features of common neighbor in social network. Its formal description is as follows:

$$Division(v, Su) \rightarrow \{v_1, Su_1\} \cup \{v_2, Su_2\} \tag{4}$$

Where v is the privacy protection demand node whose S.v=H or M. Node v is called the parent node;

Su is the 1-degree neighbor subgraph contained in the social network graph. In other words, Su is a structure subgraph which is composed of node v and its all neighbor nodes;

v_1, v_2 are two new nodes formed by the parent node division. They inherit the attribute information and social relations of node v according to the common neighbor similarity . The v_1, v_2 are called the child nodes;

Su_1, Su_2, respectively is the new neighbor subgraph of node v_1 and v_2.

3. Node division Anonymity Algorithms targeting improve dada utility

In this section, we describe the detailed implementing steps of the privacy-preserving algorithm based on node division we proposed according to the above research. And we name it as (d, k)-u anonymity algorithm, d means node division, k means k-degree anonymity, u means dada utility.

Algorithm (d, k)-u anonymity.
Input: G(V,E,A,S), k Output: (d, k)-u anonymity graph G* 1 for(i=1;i<=|V|;i++) 2 { $S.v_i$=sensitivity(v_i); 3 if ($S.v_i$=M or H) then 4 SV←v_i; } 5 if (|SV|≠0) then 6 for all v_t in SV 7 { v_{t-1}, v_{t-2}←new node(V_t) 8 for each social edge E_i of v_t 13 Distribute v_{t-1}, v_{t-2} by common neighbor number; 14 for each attribute A_i of v_t 15 Distribute v_{t-1}, v_{t-2} by attribute similarity;} 16 Return G*;

Step 0 initialization, line 2. Measure $S(A_i)$ by function sensitivity(x).

Step 1 create privacy-preserving requirement vertices set SV, line 1-4. The time complexity is O(n).

Step 2 node division, line 5-15. Assign edges and attributes of parent node into child nodes. Its time complexity is $O(n^2)$.

So the time complexity of (d, k)-u anonymity algorithm is $O(n^2)$.

4. Experiments

In this section, we report the empirical results that we conducte to evaluate the performance of our proposed (d, k)-u anonymity algorithm. All of the experiments have been implemented using MATLAB 2010a and software Gephi. The experiments were conducted on a PC having an Intel Duo 2.13GHz processor and 2GB RAM with Windows7.

4.1 Data Sets

We ues a real data sets from the musical "Les Miserables" as the experiment data. The data set contains of character and relationships between them, named Les Miserables. The Fig.3 showed the social network graph visually.

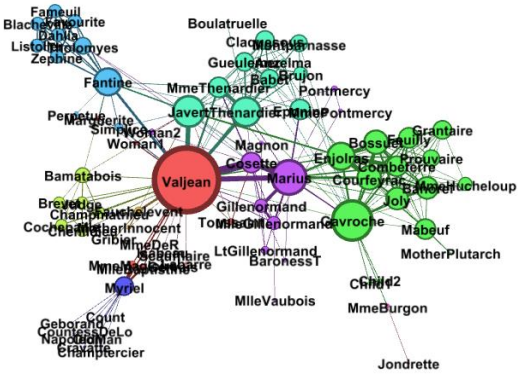

Fig.3 Les Miserables.gexf

4.2 Results and Analysis

This section compares our algorithm (d, k)-u anonymity with the existing k-degree anonymous algorithm [9] which based on node delection operations. The evaluation results are showed by running time, and data utility.

4.2.1 Running time

The Fig.4 presents the running time of our (d, k)-u algorithm and k-degree algorithm when k increases in dataset respectively. From the figures, we can observe that the largest running time is less than 30s for both algorithms. But our algorithm works much better.

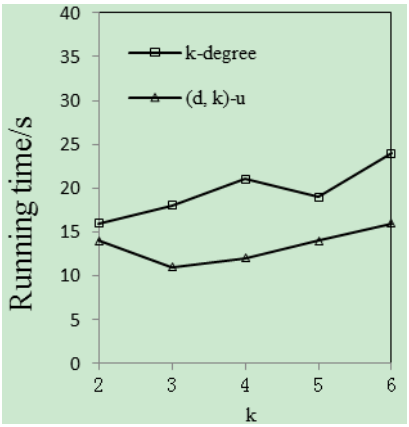

Fig.4 Les Miserables dataset: running time for different k

4.2.2 Data utility

In this part we examined how well the published graph represents the original graph with three topological structure indicators degree, APL and CC.

The Table 1 shows the difference of graph properties between original and anonymity graph by two algorithms. Apparently, the (d, k)-u anonymity method made the social network graph properties be closer to the original graph. So the algorithm we proposed may reduce the published dada utility loss effectively.

Table 1 Les Miserables dataset: data utility with k=3

structure / Graph	degree	APL	CC
Original	6.597	2.641	0.736
k-degree anonymity	5.737	2.842	0.665
(d, k)-u anonymity	6.513	2.717	0.72

5. Conclusion

In this paper, we proposed a proper privacy-preserving anonymity algorithm for social network data publishing. We used a novel anonymity operation named node division and divided attribute values according to different sensitivity aiming to reduce data utility less. Experimental evaluation on real dataset shows our approach outperforms the existing approaches in terms of the utility with the same privacy preserving.

Acknowledgements

This work is partially supported by the Natural Science Foundation of China (NO.61402266), Social Science Foundation of China (14BTQ049).

References

[1] L.A. Dunning, R. Kresman: Privacy preserving data sharing with anonymous id assignment [J]. IEEE Transactions on Information Forensics and Security, vol. 8(2013) No.2, p.402-413.

[2] X.Y. Liu, B. Wang, X.C. Yang: Survey on Privacy Preserving Techniques for Publishing Social Network Data[J], Journal of Software, vol. 25(2014) No.3, p.576-590.
[3] M. Ninggal, J. Abawajy: Utility-aware social network graph anonymization[J], Journal of Network & Computer Applications, vol.56(2015), p.137-148.
[4] M.X. Yuan, L. Chen, S.Y Philip, et al. Protecting sensitive labels in social network data anonymization [J], IEEE Transactions on Knowledge and Data Engineering, vol. 25(2013) No.3, p.633-647.
[5] K. Liu, E. Terzi: Towards identity anonymization on graphs [C], Proceedings of the 2008 ACM SIGMOD international conference on Management of data, (ACM, 2008), p. 93-106.
[6] L. Sweeney: k-anonymity: a model for protecting privacy, International Journal on Uncertainty, Fuzziness and Knowledge-based Systems, vol.10(2002) No.5, p.557–570.
[7] P. Silvia, P.A. Javier, F. Jordi, et al. On content-based recommendation and user privacy in social-tagging systems [J], Computer Standards & Interfaces, vol.41(2015), p.17-27.
[8] Y. Li, Q. Yan, et al. Privacy leakage analysis in online social networks[J], Computers & Security, vol.49(2015), p.239-254.
[9] D. Mohapatra, M.R. Patra: k-degree Closeness Anonymity: A Centrality Measure Based Approach for Network Anonymization[M], Distributed Computing and Internet Technology, Springer International Publishing, 2015,P.299-310.
[10] Y. Wang, L. Xie, B. Zheng: Utility-oriented K-anonymization on social networks[C], Database Systems for Advanced Applications, Springer Berlin Heidelberg, 2011, p.78-92.