International Journal of Innovative Research in Advanced Engineering (IJIRAE) ISSN: 2349-2163
Volume 1 Issue 8 (September 2014) www.ijirae.com

Distributed Data Anonymization with Hiding Sensitive Node Labels

C.EMELDA
Research Scholar,
PG and Research Department of Computer Science,
Nehru Memorial College, Putthanampatti,
Bharathidasan University,Trichy

R.JAYA
Assistant Professor,
PG and Research Department of Computer Science
Nehru Memorial College, Putthanampatti,
Bharathidasan University, Trichy

Abstract— Recently, people share their information via social platforms such as Face book and Twitter in their daily life. Social networks on the Internet can be regarded as a microcosm of the real world and worth being analyzed. Since the data in social networks can be private and sensitive, privacy preservation in social networks has been a focused study. Previous works develop anonymization methods for a single social network represented by a single graph, which are not enough for the analysis on the evolution of the social network. In this paper I implement the data anonymization protection model in a distributed environment, where different publishers publish their data independently and their data are overlapping. And also this work will extend to individual identifications which use nodes edge label information in graphs. The problem is defined as PSNM k- Anonymity (Protecting Sensitive Nodes Model). I conduct extensive experiments to evaluate the effectiveness of the proposed technique.

Keywords— PSNM k- Anonymity, privacy preservation, nodes edge label editing.

I. INTRODUCTION

Recently people share their information and participate in various activities on the Internet via social platforms such as Google+, Facebook and Twitter. The data in social networks are worth being analyzed in social science research since they can reflect the real social activities. Since the data in social networks include personal information and interactions, which can be private and sensitive, there is a need to anonymize the data to protect user's privacy before their release for some analysis.

A social network can be represented by a graph consisting of nodes and edges between the nodes. The nodes are used to represent users while the edges represent the interactions between the users. Recently, much work has been done on anonymizing tabular micro-data. A variety of privacy models as well as anonymization algorithms has been developed (e.g., k-anonymity, l-diversity, t-closeness). In tabular micro-data, some of the non-sensitive attributes, called quasi identifiers, can be used to re-identify individuals and their sensitive attributes. When publishing social network data, graph structures are also published with corresponding social relationships. As a result, it may be exploited as a new means to compromise privacy.

In this paper consider the privacy preservation problem in a distributed manner. A set of social network graphs are anonymized to a sequence of sanitized graphs to be released. Study the problem of protecting a nodes identity by preventing attacks from an adversary who is armed with both the node degree and edge labels information. An electronic copy can be downloaded from the Journal website. For questions on paper guidelines, please contact the journal publications committee as indicated on the journal website. Information about final paper submission is available from the conference website.

II. RELATED WORKS

A big issue in publishing social network data is that may expose individuals' private or sensitive information (e.g., salary, disease, connection to a specific group of people). A simple method used in the past, to address this issue, is to replace the true identity of the individuals in a network with random pseudo-identifiers before releasing the data. However, Backstrom et al. [1] and Narayanan et al. [2] have shown that this simple approach cannot prevent the identification of individuals when adversaries have background knowledge about the network. To protect people's privacy from different types of attacks from adversaries, various data anonymization methodologies (e.g., k-anonymity [3], l-diversity [4], t-closeness [5]) have been proposed. In tabular microdata, some of the nonsensitive attributes, called quasi identifiers, can be used to reidentify individuals and their sensitive attributes. When publishing social network data, graph structures are also published with corresponding social relationships. As a result, it may be exploited as a new means to compromise privacy.

A structure attack refers to an attack that uses the structure information, such as the degree and the subgraph of a node, to identify the node. To prevent structure attacks, a published graph should satisfy k-anonymity. The goal is to publish a social graph, which always has at least k candidates in different attack scenarios in order to protect privacy.

Liu and Terzi [6] did pioneer work in this direction that defined a k-degree anonymity model to prevent degree attacks (Attacks use the degree of a node). A graph is k-degree anonymous if and only if for any node in this graph, there exist at least k - 1 other node with the same degree.

Current approaches for protecting graph privacy can be classified into two categories: clustering and edge editing.

The edge editing- based model is to add or delete edges to make the graph satisfy certain properties according to the privacy requirements. Most edge-editing-based graph protection models implement k-anonymity of nodes on different background knowledge of the attacker.

Liu and Terzi [6] defined and implemented k-degree-anonymous model on network structure that is for published network, for any node, there exists at least other k-1 nodes have the same degree as this node.

Zhou and Pei [7] considered k-neighborhood anonymous model: for every node, there exist at least other k-1 nodes sharing isomorphic neighborhoods.

Zou et al. [8] proposed a k-Automorphism protection model: A graph is k-Automorphism if and only if for every node there exist at least k - 1 other nodes do not have any structure difference with it.

Cheng et al. [9] designed a k-isomorphism model to protect both nodes and links: a graph is k-isomorphism if this graph consists of k disjoint isomorphic sub-graphs. The sensitive attributes of nodes are protected by anatomy model [10] in a kisomorphism graph.

Ying and Wu [11] proposed a protection model which randomly changes the edges in the graph.

Hay et al. [12] proposed a heuristic clustering algorithm to prevent privacy leakage using vertex refinement, subgraph, and hub-print attacks. Zheleva and Getoor [13] developed a clustering method to prevent the sensitive link leakage. Cormode et al. [14] introduced (k,l)- clusterings for bipartite graphs and interaction graphs, respectively.

After they are preserve important graph properties, such as distances between nodes by adding certain “noise” nodes into a graph. This idea is based on the following key observation. Most social networks satisfy the Power Law distribution [15] i.e., there exist a large number of low degree vertices in the graph which could be used to hide added noise nodes from being reidentified. By carefully inserting noise nodes, some graph properties could be better preserved than a pure edge-editing method.

This paper proposes a better idea to important node degree anonymization the values (node lables) are anonymized using fuzzy concept.

III. METHODOLOGY

This paper proposes a new anonymity model for social networks in a distributed manner. Different Data publisher publish their data independently and their data are overlapping. In a distributed environment, although the data published by each publisher satisfy certain privacy requirements. Here the new method is designed to help these publishers publish a unified data together to guarantee the privacy. The social network contains both nodes label and edges. The previous work focuses only the edges. It does not consider the node labels. This work focuses both node labels and edges. This work contains two phase Node Degree Anonymization and Edge Label Anonymization. After node degree anonymization the values (node lables) are anonymized using fuzzy concept.

The first phase to achieve anonymity of a graph is to perform edge degree anonymization. The purpose of this step is to add as few edges as possible to the graph G to create graph G’ such that each node in G’ has the same degree as at least K-1 other nodes.

In this paper the node grouping based degree anonymization algorithm is proposed. In this algorithm, the sorted nodes are first partitioned into different groups such that each group consists of K nodes. Then, all the groups are traversed and anonymized. In this step, when anonymizing the nodes in one group, the degrees of each node should be increased to be the same as the first node's degree d_0. Once the degrees of a group's nodes are the same, this group is denoted as anonymized, and all its nodes are marked as “anonymized". The major operation in anonymizing a group's node is to add edges for every node v to make its degree the same as d_0. To add an edge for v, the algorithm chooses another unanonymized node. In case there are insufficient numbers of unanonymized nodes to create edges for v, anonymized nodes are randomly selected to create edges for v. accordingly these anonymized nodes' groups are marked as unanonymized.

A. ALGORITHM

Algorithm 1: Node Degree Anonymization

S_d = The sequence with sorted node degrees for the original graph G. $S_d = (S_d(v_1), , , S_d(v_n))$

Input: Nodes

Output: Degree Anonymized

1. Partition all nodes into groups and mark each group as unanonymized
2. for each unanonymized group g
   d_0 = degree of the first node in this group
   For each node v in graph g such that Sd(v)<d_0
      i. If there are d_0 −Sd(v) unanonymized nodes in V , randomly choose
      d_0−Sd(v) nodes and create edges between v and these nodes
      ii. Otherwise (i.e., there are insufficient number of unanonymized nodes to create edges for v)
         - Randomly select anonymized nodes v' to create edges between v and these nodes
         - Mark the group that each v' belongs to as unanonymized
   Mark every node in g as anonymized

The second phase is edge label anonymization. In this phase the edge labels are anonymized using fuzzy concept. The following algorithm is used for edge data anonymization

Algorithm 2: Data Anonymization

Input: Data Set D

***Output*:** Anonymized datasets AD

1. Read input Data Set D
2. Convert Categorical Attribute into Numerical Attribute
3. For each attribute apply fuzzy concept

$$F(x) = \begin{cases} 2*[(x-a)/(b-a)]^2, & a \le x \le (a+b)/2 \\ 1-2*[(x-b)/(b-a)]^2, & ((a+b)/2) \le x \le b \\ 1, & x \ge b \\ 0, & otherwise \end{cases}$$

Here

a= minimum value
b= maximum value of the particular column
x= value in that column

4. Transform the fuzzy values into Low, Medium, High and Very High
5. Return Anonymized dataset

B. EXPERIMENTAL RESULT

This section presents the experimental results on the performance of our proposed techniques. The approaches are implemented using JAVA. All the experiments were run on a Windows 7 with an Intel Pentium(R) CPU P6200 (@2.13GHz) and 2GB RAM.

The adult data is used for our experiments. The Data set contains the following list of attributes: Node ID, Age, Sex, Work Class, Education, Occupation, Marital, Race, Salary and Number of connected edges.

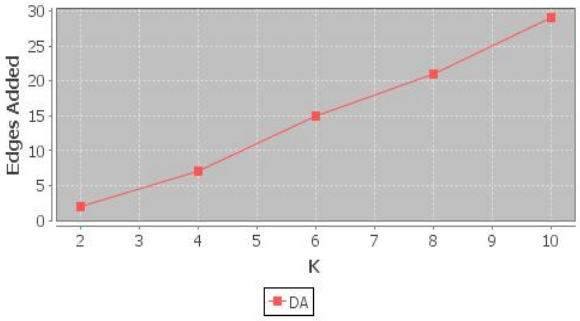

Figure 1 No of Edges Added

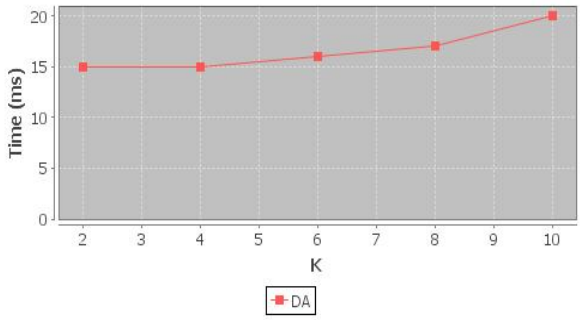

Figure 2 Execution Time

The set of experiments to test the degree anonymization techniques are shown in Figure 1 and 2. Figure 1 shows the number of new edges added to original data set. Figure 2 shows the execution time of degree anonymization.

IV. CONCLUSIONS

In this paper, we study the problem of achieving K-anonymity of social networks containing rich information to prevent attacks utilizing both node degrees and edge labels information. The major challenge in solving this problem is because of the correlation of edge labels. To address this challenge, we present a two-phase solution framework, which anonymizes edge degrees and edge labels in two phases. The first phase, degree anonymization and the second phase label anonymization. We theoretical analyze the optimality and running time complexity for the techniques in different stages of this framework.

ACKNOWLEDGMENT

I am very glad to express my deep sense of gratitude and profound thanks to my research advisor **Mrs. R. JAYA, M.Sc., M.Phil.,** Department of Computer Science, Nehru memorial college (Autonomous), Puthanampatti, Tiruchirapalli – Dt, for required guidance and encouragement at every stage of this research work. I feel proud in sharing this success with staff members and friends who helped me with their timely help and co-operation in successful completion of the research work.

REFERENCES

[1]. L. Backstrom, C. Dwork, and J. M. Kleinberg. Wherefore art thou r3579x?: anonymized social networks, hidden patterns, and structural steganography. In WWW, pages 181{190, 2007.
[2]. A. Narayanan and V. Shmatikov. De-anonymizing social networks. In IEEE Symposium on Security and Privacy, pages 173{187, 2009
[3] L. Sweeney, "K-Anonymity: A Model for Protecting Privacy," Int'l J. Uncertain. Fuzziness Knowledge-Based Systems, vol. 10, pp. 557- 570, 2002
[4] A. Machanavajjhala, D. Kifer, J. Gehrke, and M. Venkitasubramaniam, "L-Diversity: Privacy Beyond K-Anonymity," ACM Trans. Knowledge Discovery Data, vol. 1, article 3, Mar. 2007
[5] N. Li and T. Li, "T-Closeness: Privacy Beyond K-Anonymity and L-Diversity," Proc. IEEE 23rd Int'l Conf. Data Eng. (ICDE '07), pp. 106-115, 2007.
[6] K. Liu and E. Terzi, "Towards Identity Anonymization on Graphs," SIGMOD '08: Proc. ACM SIGMOD Int'l Conf. Management of Data, pp. 93-106, 2008
[7] B. Zhou and J. Pei, "Preserving Privacy in Social Networks Against Neighborhood Attacks," Proc. IEEE 24th Int'l Conf. Data Eng. (ICDE '08), pp. 506-515, 2008.
[8]. L. Zou, L. Chen, and M.T. O ¨ zsu, "K-Automorphism: A General Framework for Privacy Preserving Network Publication," Proc. VLDB Endowment, vol. 2, pp. 946-957, 2009
[9] J. Cheng, A.W.-c. Fu, and J. Liu, "K-Isomorphism: Privacy Preserving Network Publication against Structural Attacks," Proc. Int'l Conf. Management of Data, pp. 459-470, 2010
[10] X. Xiao and Y. Tao, "Anatomy: Simple and Effective Privacy Preservation," Proc. 32nd Int'l Conf. Very Large Databases (VLDB '06), pp. 139-150, 2006.
[11] X. Ying and X. Wu, "Randomizing Social Networks: A Spectrum Preserving Approach," Proc. Eighth SIAM Conf. Data Mining (SDM '08), 2008
[12] M. Hay, G. Miklau, D. Jensen, D. Towsley, and P. Weis, "Resisting Structural Re-Identification in Anonymized Social Networks," Proc. VLDB Endowment, vol. 1, pp. 102-114, 2008

[13] E. Zheleva and L. Getoor, "Preserving the Privacy of Sensitive Relationships in Graph Data," Proc. First SIGKDD Int'l Workshop Privacy, Security, and Trust in KDD (PinKDD '07), pp. 153-171, 2007
[14] G. Cormode, D. Srivastava, T. Yu, and Q. Zhang, "Anonymizing Bipartite Graph Data Using Safe Groupings," Proc. VLDB Endowment, vol. 1, pp. 833-844, 2008
[15] A.-L. Baraba´si and R. Albert, "Emergence of Scaling in Random Networks," Science, vol. 286, pp. 509-512, 1999.
[16] Mingxuan Yuan,Lei Chen, Member, IEEE, Philip S.Yu, Fellow, IEEE, and Ting Yu, "Protecting Sensitive Labels in Social Network Data Anonymization" IEEE Transactions on Knowledge and Engineering, Vol.25, No.3 March 2013.

