Protecting Sensitive Labels in Social Network Data

Navnath S. Bagal
Post Graduate Student,
Department of Computer Engg.
PVPIT Bavdhan Pune-21

Navnath D. Kale
Assistant Professor,
Department of Computer Engg.
PVPIT Bavdhan Pune-21

ABSTRACT
Privacy is one of the major concerns when publishing or sharing social network data for social science research and business analysis. Recently, researchers have developed privacy models similar to k-anonymity to prevent node re-identification through structure information. However, even when these privacy models are enforced, an attacker may still be able to infer one's private information if a group of nodes largely share the same sensitive labels (i.e., attributes). In other words, the label-node relationship is not well protected by pure structure anonymization methods. Furthermore, existing approaches, which rely on edge editing or node clustering, may significantly alter key graph properties. In this paper, we define a k-degree-l-diversity anonymity model that considers the protection of structural information as well as sensitive labels of individuals. We had seen a novel anonymization methodology based on adding noise nodes. We implemented that algorithm by adding noise nodes into the original graph with the consideration of introducing the least distortion to graph properties. We here propose novel approach to reduce number of noise node so that decrease the complexity within networks. We implement this protection model in a distributed environment, where different publishers publish their data independently Most importantly, we provide a rigorous analysis of the theoretical bounds on the number of noise nodes added and their impacts on an important graph property. We conduct extensive experiments to evaluate the effectiveness of the proposed technique.

General Terms

Protecting sensitive information using KDLD technique and graph increase/decrease, label setting, sequence generation algorithms.

Keywords

Privacy, Online Social Network, Privacy protecting in SN, Sensitive information

1. INTRODUCTION
Protecting the privacy of personal information is one of the biggest challenges facing website developers, especially social network providers. Several researchers have discussed the issue of privacy. In today's internet determined the people we have witnessed the rapid growth of online social networking sites (OSN) as well as their integration into our everyday life. OSN such as Facebook (FB), Twitter, LinkedIn, Myspace etc. now represent a fundamental shift in the way that we communicate in our personal and working live. With the sharing nature of OSN's and the sites' control of posted information and personal relationships, concerns have developed regarding trust and privacy issues within social networking. Mainly, the data may contain sensitive information about individuals that cannot be disclosed without compromising their confidentiality. This paper we use AES algorithm to encrypt sensitive attributes and attribute names. We used unique token (key) per user, therefore, prevents potential leaks of sensitive labels and information associated with them. Because it will be publish in encrypted format.
We consider a graph model in every vertex of graph is linked with sensitive labels or private information. We develop a new algorithm (heuristic search) by adding noise nodes into the original graph without change original graph drastically, and provide security of each user & its sensitive data.

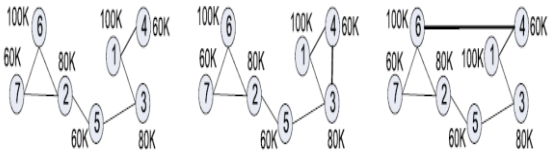


Fig. a) Original graph SN b)2-degree anonymous graph c) 2degree 2-diversity graph SN

Fig. 1a shows an example of a possible structure attack using degree collect the information. If an adversary knows that one person has three friends in a graph, he can know that node 2 is that person and the related attributes of node 2 are revealed. K-degree anonymity can be used to prevent such structured attacks in SN. However, in many applications in, a social network where each node has sensitive attributes should be published. For example, a graph may contain the user salaries which are sensitive. In this case, k-degree alone is not sufficient to prevent the inference of sensitive attributes of individuals. Fig. 1b shows a graph that satisfies 2-degree anonymity but node labels are not consider in a graph. In it, nodes 2 and 3 have the same degree 3, but they both have the label "80K." If an attacker knows someone has three friends in the social networks, he can conclude that this person's salary is 80K without exactly re-identify the node. Therefore, when sensitive labels are considered, the l-diversity should be adopted for graphs. Again, the l-diversity concept here has the same meaning as that defined over tabular data.

2. LITERATURE REVIEW
Online social Networks have always been an important component of our daily life, but currently that more and more people are connected to the Internet, and their online counterpart is satisfying an increasingly vital role. Consider a graph model where each vertex in the graph is associated with as the sensitive label or (private information). According to survey privacy related issues in social networking is very important. Since this work explores the Preserving privacy in publishing social network data becomes an important concern. With some local knowledge about individuals in a social network, an adversary may attack the privacy of some victims

easily. Unfortunately, most of the previous studies on privacy preservation data publishing can deal with relational data only, and cannot be applied to social network data. In this paper, we take an initiative towards preserving privacy in social network data. Specifically, we identify an essential type of privacy attacks: neighborhood attacks. If an adversary has some knowledge about the neighbors of a target victim and the relationship among the neighbors, the victim may be re-identified from a social network even if the victim's identity is preserved using the conventional anonymization techniques. To protect privacy against neighborhood attacks, we extend the conventional k-anonymity and l-diversity models from relational data to social network data. We show that the problems of computing optimal k-anonymous and l-diverse social networks are NP-hard. We develop practical solutions to the problems. The empirical study indicates that the anonymized social network data by our methods can still be used to answer aggregate network queries with high accuracy. The increasing popularity of social networks has initiated a fertile research area in information extraction and data mining. Although such analysis can facilitate better understanding of sociological, behavioral, and other interesting phenomena, there is growing concern about personal privacy being breached, thereby requiring effective anonymization techniques.

2.1 SN Anonymization via edge editing

Bruce Kapron, Gautam Srivastava, S. Venkatesh - Provide a framework to show hardness results for different variants of social network anonymization using a common approach. Start by showing that k-label sequence anonymity of arbitrary labeled graphs is hard, and use this result to prove NP-hardness results for many other recently proposed notions of graph anonymization. Secondly, we present interesting algorithms and hardness for bipartite graphs. For unlabeled bipartite graphs, we show k-degree anonymity is in P for all k_2. For labeled bipartite graphs, we show that k-label sequence anonymity is in P for k = 2 but it is NP-hard for k_3.

2.2 Anonymization technique for privacy

Bin Zhou, Jian Pei, Wo-Shun Luk- Privacy preserving publishing of social network data becomes a more and more important concern. In this paper, we present a brief yet systematic review of the existing anonymization techniques for privacy preserving publishing of social network data. We identify the new challenges in privacy preserving publishing of social network data comparing to the extensively studied relational case, and examine the possible problem formulation in three important dimensions: privacy, background knowledge, and data utility. We survey the existing anonymization methods for privacy preservation in two categories: clustering-based approaches and graph modification approaches.

3. PROBLEM DISCRIPTION

The publication of social network data entails a privacy threat for their users. Sensitive information about users of the social networks should be protected. The challenge is to devise methods to publish social network data in a form that affords utility without compromising privacy. Previous research has pro-posed various privacy models with the corresponding protection mechanisms that prevent both inadvertent private information leakage and attacks by malicious adversaries. The social networks are modeled as graphs in which users are nodes and features are labels. Labels are denoted either as sensitive or as non-sensitive. Each node in the graph represents a use. Main challenge is how it works in distributed environment.

3.1 Objective

To develop a new technique to provide privacy and security of social network data in distributed environment with the help of graph property.

The objectives of project are as follows:

- We can publish the Non sensitive data to every-one in social Network.
- Add minimum no of noise and improve anonymization technique.
- Increase and decrease the graph with edge editing.
- Assign sensitive label to noise node.
- Protecting sensitive data of each individual user.
- Security in distributed architecture.

4. IMPLEMENTATION DETAILS

Anonymization is a clustering problem one or more nodes are connected each other in various graph in social network and sharing information and resources in social networking business as well as personal relations.

4.1 Mathematical Model

Social Network Graph: a social network graph is a four tuple G (V, E, σ, ≺) where V is a set of vertices, and each vertex represents a node in the social network. E ⊆ V × V is the set of edges between vertices, _ is a set of labels that vertices have.≺: V -> σ maps vertices to their labels.

Since each noise node connects with at least one noise edge, "Low Overhead" also limits the number of noise nodes that can be added. The social distance between two nodes u, v is the shortest path length between u and v in the original graph. The social distance between all node pairs of a graph is measured by Average shortest path length (APL). APL is a concept in network topology that is defined as the average of distances between all pair's of nodes. It is a measure of the efficiency of information or mass transport on a network. Some queries like "the nearest node for a group of nodes" are related to APL. The APL of a graph G is

$$APL_G = 2/N(N-1) \sum_{\forall n_i n_j \in G} d(n_i, n_j) \quad \text{.....(I)}$$

Where, d (ni, nj) is the length of the shortest path between nodes ni and nj, N is the number of nodes in the graph. We design a two-step algorithm to generate the KDLD graph which tries to preserve the above two key properties. In the first step, we compute a target degree for each node so that it makes the original graph satisfy KDLD constraint with the minimum sum of degree change. Clearly, smaller degree change needs fewer noise edges to implement the change. In the second step, we change each node's degree to its target degree by adding noise edges/nodes. We utilize the noise nodes to make the change of APL as small as possible. Our proposed two step algorithm considers both the "Low Overhead" and the "Preserve Social Distance" requirements.

4.1.1 Definition 3:

Given a graph G, its sensitive degree sequence is a sequence of n triples: [P[1], . . . , P[n]] where P[1]:d >P[2]

d>……> P[n]: d, P[i] is a triple (id, d, s) at position i in P, d is the degree, and s is the sensitive label associated with node id.

4.1.2 Definition 4:

KDLD sequence: A sensitive degree sequence P is a KDLD sequence if P satisfies the following constraint: P can be divided into a group of subsequences: [P[1], . . . , P[n]], : [P[1], . . . , P[n]]such that for any subsequence

[P[1], . . . , P[n]]satisfies three constraints:1) All the elements in Px share the same degree(P[ix]: P[ix]1:d = _ _ _ = P[ix]:d);2) Px has size at least k(ixþ1 - ix + 1 >=k);3) Px's label set {P[t] sji x< t < ixþ1} have at least l distinct values.

- We first generate KDLD sequence. In our dissertation using definition using 3 & 4 KDLD sequence is calculated.
- To generate KDLD sequence triples p in groups all the nodes in one group. We have same degree.
- We then used two algorithm that put nodes with similar degree to reduce degree changes.

We then calculate node degree of nodes one by one such that k degree, L diversity is maintain.

4.2 Algorithmic Strategy

To generate a KDLD sequence is calculate, the triples P in P should be divided into groups in a graph. All the corresponding nodes in the same group shall be adjusted to have the same degree in a graph.

4.2.1 Algorithm two: - Steps

We are using the following algorithm to construct the published graph which preserves the APL. The algorithm contains five steps are as follows: Adding, editing and deleting extra noise nodes in original graph.

Step 1: Neighborhood_Edge_Editing ()

We add or delete some edges if the corresponding edge-editing operation follows the neighbourhood rule. By doing this, the sensitive degree sequence P of original graph G is closer to P new in case AP L is preserved;

Step 2: Adding_Node_Decrease_Degree ()

For any node whose degree is larger han its target degree in P new , we decrease its degree to the target degree by making using of noise nodes;

Step 3: Adding_Node_Increase_Degree ()

6. EXPERIMENTAL RESULT

The effectiveness of our anonymization algorithm, we compare our work with two pure edge-editing graph construction algorithms: adding edges and graph construction algorithm. We generate the KDLD graph for each data set using the K-L-BASED sensitive degree sequence generation algorithm. Note here we use different graph construction algorithms to generate anonym zed graphs for the same KDLD sequence. In this experiment we exam our algorithm on three real data set – Arnet (Nodes and Edges), Cora data

For any node whose degree is smaller than its target degree in P new, we increase its degree to the target degree by making using of noise nodes;

Step 4: New_Node_Degree_Setting ()

For any noise node, if its degree does not appear in P new, we do some adjustment to make it has a degree in P new. Then, the noise nodes are added into the same degree groups in P new;

Step 5: New_Node_Label_Setting ()

We assigning sensitive labels to noise nodes to make sure all the same degree groups still satisfy the requirement of the distinct l -diversity. It is obvious that after Step 4 and Step 5, the sensitive degree sequence P' of the published graph G0 is a KDLD sequence. In Steps 2, 3, and 4, we carefully connect the noise nodes into the graph to make the change of AP L as less as possible in a graph.

5. SYSTEM ARCHITUCTURE

Fig 5.1 shows the system architecture Anonymizaion techniques uses, distance between nodes & edges are measured, anonymization algorithm (Noise node adding) are for privacy preserving of graph.

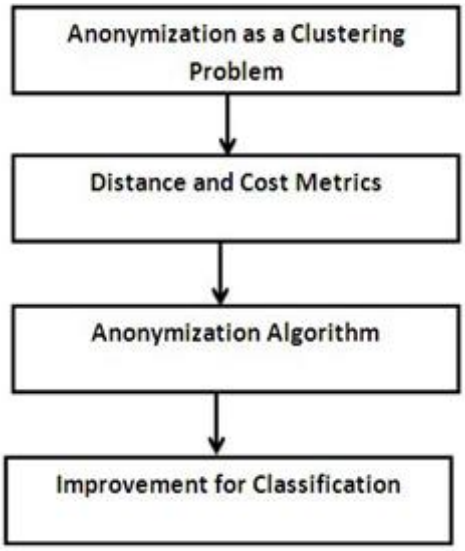


Figure 5.1 System Architecture

In system architecture following components are important.

- Anonymization as a clustering problem.
- Distance and cost of graph are measured.
- Applying Edge editing algorithm.
- Assigning Sensitive labels and sequence generation.
- Improve anonymization technique.
- Protecting privacy in distributed environment.

set (Nodes and Edges), and DBLP data set (Nodes and Edges) Details of these data set can be found in online supplement material and results. With the help of noise node adding algorithm

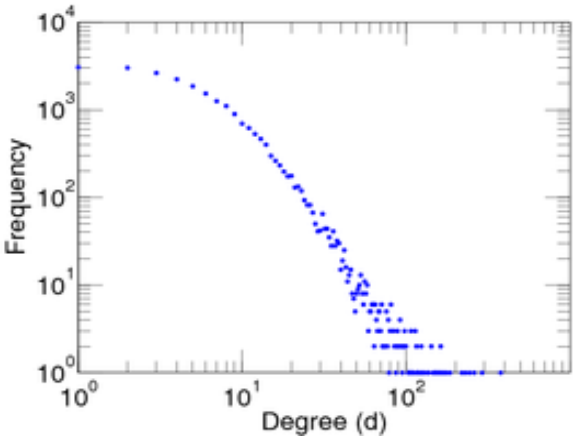
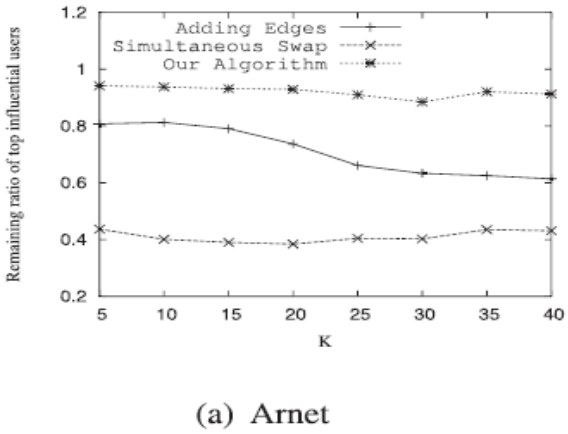


Figure.6.1 Cora Dataset



(a) Arnet

Figure.6.2 Arnet Dataset

7. TEST CASES

The quality of the published graph, we also test several other aspects of our algorithm. Algorithm efficiency and noise nose adding ratio.

7.1 Algorithm Efficiency

We record the running time of our algorithm for different k In Fig. 7.1 from the result we can see our algorithm is very Efficient,

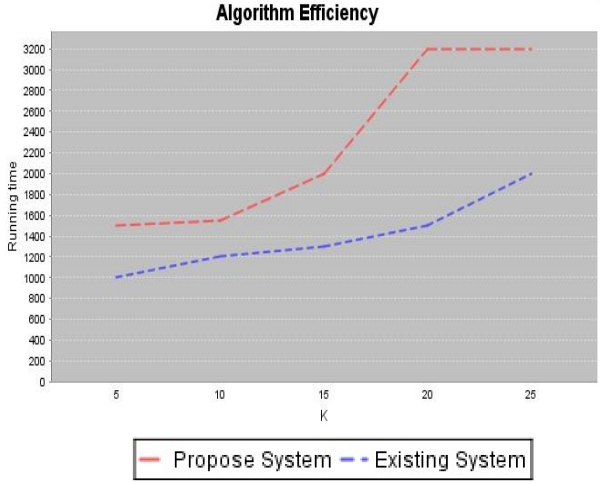


Figure 7.1 Algorithm efficiency

7.2 Percentage of Noise Nodes

Fig.7.2 shows the percentage of noise nodes added by our algorithm with different ks. Our method only put a small "burden" to achieve a much better effect comparing with pure edge-editing algorithms.

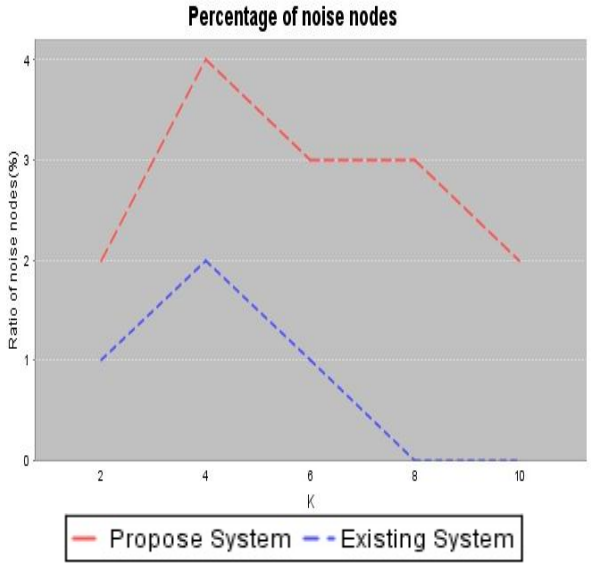


Figure 7.2 Percentage of noise node

8. CONCLUSION

We propose a k-degree-l-diversity model for privacy preserving social network data publishing. We implement distinct K-degree, l-diversity and Anonymization. We design an algorithm to preserving privacy of user on social network. We are using heuristic search strategy that will search the input phase with minimum overhead. With give approximate answer within polynomial time. We give a rigorous analysis of the theoretical bounds on the minimum number of noise nodes added. Extensive experimental results demonstrate that the add minimum noise node AES algorithms and heuristic strategy can achieve a better result than the previous work using edge editing only and noise node adding attractive direction to study clever algorithms which can reduce the reduction of noise nodes with anonymization and diversity. Privacy is key matter when sharing social network data for organization and personal. It is necessary of today's large use of social network to provide privacy and security of private information. We present new technique that will reduce noise nodes in our model

- Add minimum no of nodes & improve anonymization technique.
- We implementing privacy-preserving approach.

It is designed to help out these publishers publish an integrated data together to certification the security and privacy.

9. REFERENCES

[1] K. Le-Fevre, D. DeWitt, R. Ramakrishnan. Mondrian multidimensional k-anonymity In International Conference on Data Engineering 2006.

[2] M. Hay, G. Miklau, D. Jensen, D. Towsley, P. Weis, "Resisting Structural re-Identification in the Anonymized Social Networks," Proc. VLDB Endowment, vol. 1, pp. 102-114, 2008.

[3] 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. 08), 2008.

[4] Hay, Michael; Miklau, Gerome; Jensen, David; Weis, Philipp; and Srivastava, Siddharth, "Anonymizing Social Networks" (2007).Computer Science Department Faculty Publication Series paper 180.

[5] K. Le-Fevre, D. DeWitt, R. Ramakrishnan. Mondrian multidimensional k-anonymity In International Conference on Data Engineering 2006

[6] A. Meyerson and R. Williams. On the complexity of optimal k-anonymity in ACM Symposium on Principles of Database Systems 2004

[7] B.S. Hettich and C. Merz. UCI repository of machine learning databases, 1998

[8] P. Samarat,- Protecting respondent's privacy in micro data release IEEE Transactions on Knowledge and Data Engineering, 13, 2001.

[9] L. sweeney, achieving k-anonymity privacy protection using generalization and suppression. International journal on uncertainty, Fuzziness and knowledge based system, 2002.

[10] A.-L. Baraba´ si and R. Albert, "Emergence of Scaling in Random Networks," Science, vol. 286, pp. 509-512, 1999.

[11] Bruce Kapron, GautamSrivastava, S. Venkatesh -IEEE international Conference 2011, Social Network anonymization via Edge Addition.

[12] Benjamin C. M. Fung, Ke Wang, and Philip S.Yu, Fellow, IEEE Data Engineering 2007 AnonymizingClassication Data for Privacy Preservation.

[13] Ping Xiong, Tianqing Zhu management of e-Commerce and e Government (ICMeCG), 2012 Conference on An AnonymizationMethod Based on Tradeoff between Utility and Privacy for Data Publishing.

[14] Gionis A.; Tassa, T, IEEE Knowledge and data engineering 2009.K anonymization with minimal loss of information.

[15] Shapiro, S S. (SysCon) IEEE Knowledge and data engineering 2012, Situating Anonymization within a Privacy risk model.