A.Stalin Irudhaya Raj *et al*, International Journal of Computer Science and Mobile Applications,

Vol.2 Issue. 1, January- 2014, pg. 86-93 **ISSN: 2321-8363**

Securing Sensitive Information in Social Network Data Anonymization

Mr. A.Stalin Irudhaya Raj[1], Ms. N.Radhika[2]

[1]PG Student, [2]Assistant Professor, Department of Computer Science and Engineering,

PRIST University, Trichy District, India

([1] stalin.arockiasamy@yahoo.com)

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 reidentification 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, k-degree-l-diversity anonymity model that considers the protection of structural information as well as sensitive labels of individuals. A novel anonymization methodology based on adding noise nodes has proposed. New algorithm by adding noise nodes into the original graph with the consideration of introducing the least distortion to graph properties. Most importantly, completed the rigorous analysis of the theoretical bounds on the number of noise nodes added and their impacts on an important graph property. Extensive experiments used to evaluate the effectiveness of the proposed technique.

Index Terms — ***Knowledge and Data Engineering, Data mining.***

I.INTRODUCTION

With the rapid growth of social networks, such as Facebook and Linkedin, more and more researchers found that it is a great opportunity to obtain useful information from these social network data, such as the user behavior, community growth, disease spreading, etc. However, it is paramount that published social network data should not reveal private information of individuals. Thus, how to protect individual's privacy and at the same time preserve the utility of social network data becomes a challenging topic. In this paper, a graph model where each vertex in the graph is associated with a sensitive label. Recently, much work has been done on anonymizing tabular microdata. A variety of privacy models as well as anonymization algorithms have been developed. 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 did pioneer work in this direction that defined a k-degree anonymity model to prevent degree attacks (Attacks use the degree of a node).

Vol.2 Issue. 1, January- 2014, pg. 86-93 **ISSN: 2321-8363**

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. If an adversary knows that one person has three friends in the graph, he can immediately 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 structure attacks. However, in many applications, 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. A graph that satisfies 2-degree anonymity but node labels are not considered. 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 network, he can conclude that this person's salary is 80K without exactly reidentifying 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. For example, if the distinct l-diversity, for the nodes with the same degree, their associated sensitive labels must have l distinct values. For each distinct degree appearing in this graph, there exist at least two nodes. Moreover, for those nodes with the same degree, they contain at least two distinct sensitive labels. Thus, the attacker cannot reidentify a node or find the node-label relation with degree knowledge. In this paper, select the degree-attack, one of the popular attacks methods, to show design mechanisms to protect both identities and sensitive labels.

With respect to other types of attacks, such as subgraph query attacks or hub node query attacks, that the key ideas proposed in this work can be adopted to handle them as well, though more complicated extensions may be needed. Current approaches for protecting graph privacy can be classified into two categories: clustering and edge editing. Clustering is to merge a subgraph to one super node, which is unsuitable for sensitive labeled graphs, since when a group of nodes are merged into one super node, the node-label relations have been lost. Edge-editing methods keep the nodes in the original graph unchanged and only add/delete/swap edges. For example, to protect privacy, and convert it to satisfy 3-degree anonymous and 3-diversity by adding edges.

However, edge editing may largely destroy the properties of a graph. The edge editing method sometimes may change the distance properties substantially by connecting two faraway nodes together or deleting the bridge link between two communities. In the distance between nodes 6 and 12 is changed from 5 to 1 hop. This phenomenon is not preferred. Mining over these data might get the wrong conclusion about how the salaries are distributed in the society. Therefore, solely relying on edge editing may not be a good solution to preserve data utility. To address this issue, A novel idea to 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 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.

The distances between the original nodes are mostly preserved. Our privacy preserving goal is to prevent an attacker from reidentifying a user and finding the fact that a certain user has a specific sensitive value. To achieve this goal, to define a k-degree-l-diversity (KDLD) model for safely publishing a labeled graph, and then develop corresponding graph anonymization algorithms with the least distortion to the properties of the original graph, such as degrees and distances between nodes.

Analytical results to show the relationship between the number of noise nodes added and their impacts on an important graph property. Further conduct comprehensive experiments for both distinct l-diversity and recursive (c,l)-diversity to show our technique's effectiveness.

Vol.2 Issue. 1, January- 2014, pg. 86-93 **ISSN: 2321-8363**

II RELATED WORK

2.1 Project overview

To secure sensitive Information in social network data anonymization using k-degree-l-diversity anonymity model.

2.1.1 Scope of project

- Privacy is one of the major concerns when publishing or sharing social network data for social science research and business analysis.
- The label-node relationship is not well protected by pure structure anonymization methods.
- k-degree-l-diversity anonymity model that considers the protection of structural information as well as sensitive labels of individuals.
- Adding noise nodes into the original graph with the consideration of introducing the least distortion to graph properties.

2.2 Existing system

2.2.1. Edge-Editing –Based Model

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 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 considered k-neighborhood anonymous model: for every node, there exist at least other k-1 nodes sharing isomorphic neighborhoods.

2.2.2. Clustering-Based Model

Clustering-based model is to cluster “similar” nodes together to form super nodes. Each super node represents several nodes which are also called a “cluster.” Then, the links between nodes are represented as the edges between super nodes which is called “super edges.” Each super edge may represent more than one edge in the original graph. The graph that only contains super nodes and super edges are called as clustered graph.

2.2.3. Disadvantages

- Simply removing the identifiers in social networks does not guarantee privacy

2.3 Proposed system

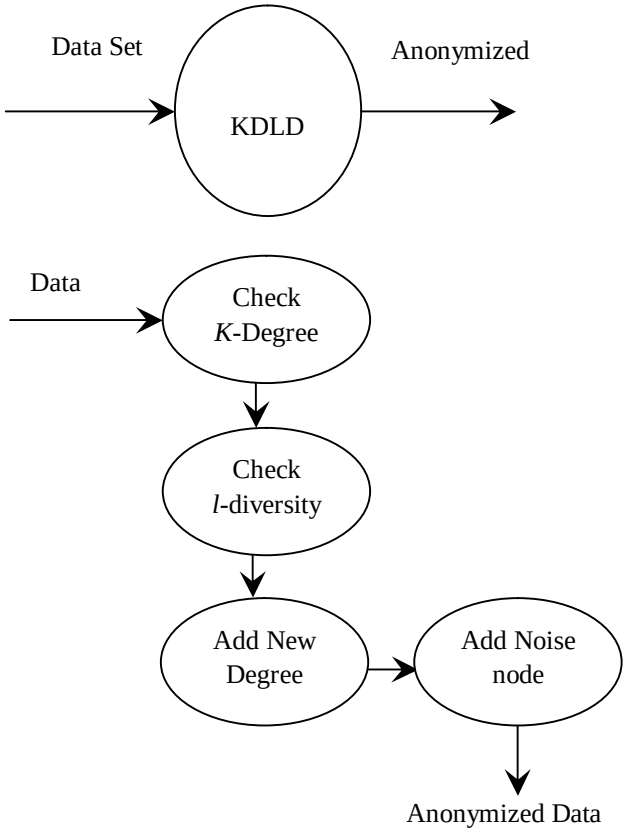
k-degree anonymity with l-diversity to prevent not only the reidentification of individual nodes but also the revelation of a sensitive attribute associated with each node. If the k-degree-l-diversity constraint satisfies create KDLD graph. A KDLD graph protects two aspects of each user when an attacker uses degree information to attack

A novel graph construction technique which makes use of noise nodes to preserve utilities of the original graph. Two key properties are considered: Add as few noise edges as possible. Change the distance between nodes as less as possible. The noise edges/nodes added should connect nodes that are close with respect to the social distance. There exist a large number of low degree vertices in the graph which could be used to hide added noise nodes from being re-identified. By carefully inserting noise nodes, some graph properties could be better preserved than a pure edge-editing method.

2.3.1 Advantages

- It helps publishers publish a unified data together to guarantee the privacy.
- Low overhead.
- Preserve social Distance.

2.4 DATA FLOW DIAGRAM



III MODULES

- Data Collection.
- Reduce Node Degree.
- Add Node Degree.
- Add Noise Node.

3.1 DATA COLLECTION

In this module the employee data is collected. Each employee has unique Id, Name and Sensitive Label Salary. Each employee links with number of other employee. Based on the employee data construct the Social Network Graph: a social network graph is a four tuple G(V, E, σ, λ), where V is a set of vertices, and each vertex represents a node in the social network. is the set of edges between vertices, σ is a set of labels that vertices have. λ: V □ σ maps vertices to their labels.

3.2 REDUCE NODE DEGREE

For any node whose degree is larger than its target degree in Pnew, decreasing its degree to the target degree by making using of noise nodes.

3.3 ADD NODE DEGREE

For any node whose degree is smaller than its target degree in Pnew, increasing its degree to the target degree by making using of noise nodes. For each vertex u in G which needs to increase its degree, to make its degree reach the target degree. First check whether there exists a node v within two hops of u, and v also needs to increase its degree. Connect n with v. Since v is within two hops of u, connecting v with n will not change the distance between u and v. After this step, if n's degree is bigger than the minimum degree in Pnew but does not appear in Pnew, recursively deleting the last created link until the degree of n equals to a degree in Pnew. Otherwise, leave n for processing and continue adding noise to u if u:d < p' u:d. By doing this, u's degree is increased to its target degree.

3.4 ADD NOISE NODE

In this module the noise node will added to the original data set. After that adding noise node add new degree for that noise node. For any noise node, if its degree does not appear in Pnew, some adjustment can happen to make it has a degree in Pnew. Then, the noise nodes are added into the same degree groups in Pnew.

IV LITERATURE REVIEW

4.1 Attacks on anonymized social networks

In this paper present both active and passive attacks on anonymized social networks, showing that both types of attacks can be used to reveal the true identities of targeted users, even from just a single anonymized copy of the network, and with a surprisingly small investment of effort by the attacker. It describe active attacks in which an adversary chooses an arbitrary set of users whose privacy it wishes to violate, creates a small number of new user accounts with edges to these targeted users, and creates a pattern of links among the new accounts with the goal of making it stand out in the anonymized graph structure. The adversary then efficiently finds these new accounts together with the targeted users in the anonymized network that is released. At a theoretical level, the creation of O(p log n) nodes by the attacker in an n-node network can begin compromising the privacy of arbitrary targeted nodes,

Vol.2 Issue. 1, January- 2014, pg. 86-93 **ISSN: 2321-8363**

with high probability for any network; in experiments, to find that on a 4.4-million-node social network, the creation of 7 nodes by an attacker (with degrees comparable to those of typical nodes in the network) can compromise the privacy of roughly 2400 edge relations on average. Moreover, experimental evidence suggests that it may be very difficult to determine whether a social network has been compromised by such an active attack.

4.2 The nature of the attacks

The social network is an n-node graph G = (V,E), representing interactions in an on-line system. Nodes correspond to user accounts, and an edge (u, v) indicates that u has communicated with v (again, consider the example Example of an e-mail or instant messaging network). The attacks become easier to carry out if the released graph data is directed; for most of the paper we will therefore consider the harder case of undirected graphs, in which we assume that the curator of the data—the agent that releases the anonymized network — eliminates the directions on the edges. The active attacks will make use of the following two types of operations. First, an individual can create a new user account on the system; this adds a new node to G. Second, a node u can decide to communicate with a node v; this adds the undirected edge (u, v) to G. The goal of the attack is to take an arbitrary set of targeted users w1, . . . ,wb , and for each pair of them, to use the anonymized copy of G to learn whether the edge (wi ,wj) in fact exists. This is the sense in which the privacy of these users will be compromised. (Other privacy compromises, such as learning the degree of a targeted user, also occur, but we focus our attention on learning about edges.)

The structure of the active attack is roughly as follows. Before the anonymized graph is produced, the attacker creates k new user accounts (for a small parameter k), and it links them together to create a subgraph H.

4.3 Anatomy: Simple and Effective Privacy Preservation:

This paper [2] presents a systematic study of the anatomy technique. First, to formalize the new methodology, based on the privacy requirement of l-diversity. Every pair of QIT and ST ensures that the sensitive value of any individual involved in the microdata can be correctly inferred by an adversary with probability at most 1/l. A larger l leads to stronger privacy protection.

4.4 K-Isomorphism: Privacy Preserving Network Publication against Structural Attacks

A social network as a simple graph, in which vertices describe entities (e.g., persons, organizations, etc.) and edges describe the relationships between entities (e.g., friends, colleagues, business partners). A vertex in the graph has an identity (e.g., name, SID) and is also associated with some information such as a set of emails. Our task is to publish the graph in such a way that, given a specific type of adversary knowledge in terms of NAGs, the NodeInfo or LinkInfo of any individual can be inferred only with a probability not higher than a pre-defined threshold, whereas the information loss of the published graph with respect to the original graph is kept small

The first half of the problem on NodeInfo security, on its own, has a simple solution. We can publish the graph structure intact with no distortion on the edges and vertices. First the content of NodeInfo for each v is screened to remove the occurrences of names or other user identifying information. The processed Node- Info of each v, I(v), is detached from vertex v. We randomly partition the vertex set V into groups of at least size k. For each group, the corresponding set of NodeInfo is published as a group. For example, if v1,….., vk form a group, then NodeInfo {I(v1),……, I(vk)} will be published as a group, breaking the linkage of the NodeInfo to each individual.

Vol.2 Issue. 1, January- 2014, pg. 86-93 **ISSN: 2321-8363**

4.5 Private Social Network Analysis: How to Assemble Pieces of a Graph Privately

TRADEOFF between privacy and accuracy

That there is a fundamental tradeoff between the privacy and accuracy of the reconstructed graph: exact private reconstruction of a graph is not always possible. This trade-off is a fundamental consequence of our adversarial assumptions and applies not only to the protocols proposed in this paper, but to any protocol that assumes the same threat model. Our threat model allows the adversary to control a subset of the subjects. If these subjects report particular patterns of relationships that "stand out" (i.e., footprints) in the graph AG, the adversary can recognize these footprints and learn part of the isomorphism that maps G to AG. The adversary can do so with knowledge only of the output of the protocol (the graph AG) and of the inputs of subjects under its control (pieces of the graph G). Note that this attack does not necessarily require the subjects controlled by the adversary to lie about their relationships: their true pattern of relationships may be easily distinguishable from the patterns of relationships of honest subjects.

V CONCLUSION

In this paper, k-degree-l-diversity model has implemented for privacy preserving social network data publishing. Implementation of both distinct l-diversity and recursive (c, l)-diversity also happened. In order to achieve the requirement of k-degree-l-diversity, a noise node adding algorithm to construct a new graph from the original graph with the constraint of introducing fewer distortions to the original graph. Rigorous analysis of the theoretical bounds on the number of noise nodes added and their impacts on an important graph property. Extensive experimental results demonstrate that the noise node adding algorithms can achieve a better result than the previous work using edge editing only. It is an interesting direction to study clever algorithms which can reduce the number of noise nodes if the noise nodes contribute to both anonymization and diversity. Another interesting direction is to consider how to implement this protection model in a distributed environment, where different publishers publish their data independently and their data are overlapping. In a distributed environment, although the data published by each publisher satisfy certain privacy requirements, an attacker can still break user's privacy by combining the data published by different publishers together. Protocols should be designed to help these publishers publish a unified data together to guarantee the privacy.

ACKNOWLEDGEMENT

I sincerely thanks to all authors in reference section. All papers in the reference section are very useful for my proposal. Their concepts and techniques are very useful for my research.

REFERENCES

[1] Lars Backstrom, Cynthia Dwork, Jon Kleinberg, "Wherefore Art Thou R3579X? Anonymized Social Networks, Hidden Patterns, and Structural Steganography".
[2] Xiaokui Xiao,Yufei Tao , "Anatomy: Simple and Effective Privacy Preservation"
[3] James Cheng, Ada Wai-Chee Fu, Jia Liu, "K-Isomorphism: Privacy Preserving Network Publication against Structural Attacks"
[4] Keith B. Frikken, Philippe Golle
"Private Social Network Analysis: How to Assemble Pieces of a Graph Privately".
[5] Srivatsava Ranjit Ganta, Shiva Kasiviswanathan,Adam Smith "Composition Attacks and Auxiliary Information in Data Privacy".
[6] Michael Hay,Gerome Miklau,David Jensen,Don Towsley,Philipp Weis "Resisting Structural Reidentification Anonymized Social Networks".

[7] Arvind Narayanan,Vitaly Shmatikov "De-anonymizing Social Networks".
[8] Bin Zhou, Jian Pei "Preserving Privacy in Social Networks Against Neighborhood Attacks".
[9] Smriti Bhagat, Graham Cormode, Balachander Krishnamurthy, Divesh Srivastava "Classbased graph anonymization for social network data".

