Protecting Sensitive Labels in Social Network Data Anonymization

Roohi Salma A.R
Vels university
Chennai-600 117

K.Dharmarajan-Scholar
Research and Development Centre
Bharathiar University
Coimbatore – 641046, India

Abstract- Privacy preservation is the major problem when sharing data's in social networks. Various Privacy models are developed to avoid node reidentification. Yet an attacker may still hack users private information, if nodes largely share the same type of sensitive labels. In this paper, we propose a scheme namely K-degree-L-diversity model useful for preserving social network data. Our anonymization methodology based on addition of noise nodes into the original social graph which reduces error rate and performs better results.

Keywords – Social networks, privacy preservation, Anonymization

I. INTRODUCTION

Data mining refers to Knowledge Discovery in Databases. The data mining process is the extraction of information from various data sets and transform to an understandable manner. There has been increasing concern about privacy of individuals when it comes to social network. Disclosing information on the web is a voluntary act of an individual, the users are unaware of who is able to access their data and how potentially their data will be used. Data privacy means freedom from unauthorized intrusion. We focus on two scenarios for privacy in social networks: privacy breaches and data anonymization. In the first scenario, the attacker is interested in knowing the information of an individual using social network data, possibly anonymized. In the second scenario, the data provider would like to share data in a social network with the researchers, but preserves the privacy of its users such that the researchers can only access the provided data alone. A very common assumption in data anonymization is that the data is described by a single table with attribute information for each of the entries. When we refer to social network we generally mean online sites such as Facebook, LinkedIn, Flickr, etc., where individuals can link to or "friend" each other, and which allows rich interactions such as joining communities or groups of their interest, or participating in group forums. These sites also often include online services which allow user to create a profile and share their preferences about items, such as tagging articles, postings and commenting on photos, and rating movies, books or music. Thus we view social network as a multi-modal graph in which there are multiple entities such as people, groups and items, where atleast one type is an individual and the links between individuals represent some sort of social tie. Each node of an individual has a profile, and profiles can have personal attributes such as age, gender etc. In order to provide security to social network users, our algorithms issue anonymized views of the graph with significantly smaller information losses and analyze their privacy and communication complexity. Formally in this, social networks are always represented as a graph, which we refer to as the social graph. The node of such a graph represents an actor and the edges represent ties between those actors.

II. ANONYMIZATION TECHNIQUES

A method that is used to publish data and protect user privacy is called anonymization. Anonymization in context with privacy can be defined as: replacing the information that can damage user privacy with the harmless information. Anonymization is one of the methods mostly used for security purpose such as preventing personal and sensitive information and decrease the success rate of attacks such as context aware spam attack and context aware phishing attack.

Anonymization techniques can be classified into four approaches: clustering, clustering with constraints, modification of graph and hybrid approach. Sometimes when the user identity is anonymized, still there are various techniques to reidentify the users. The Attackers make use of background information to undertake these kinds of attacks. If the attacker has some specific information about the victim, he is able to recognize his victim among the anonymized information.

There are three types of attacks they are: active attack, passive attack and semi-passive attack. Passive attacks are described as those in which attackers begin their work to detect the identity of nodes after anonymized information is published. In active attacks, the attacker tries to create some accounts in the social networking sites and establish some links in the network so that these links will be present in the anonymized version of the information. In semi-passive attacks, there is no new account creation but some links are established with the target user before the information is anonymized.
Various anonymization techniques have been researched to protect the social networking sites from de-anonymization and make it difficult for the hacker to re-identify an anonymized user in online social networks.

Anonymization techniques prevent node replication attacks. The goal is to publish a social graph, in order to protect privacy. Recent approaches for protecting social graph privacy are edge editing and clustering method. Edge-editing method keep the nodes unchanged and by increasing or decreasing or by swapping edges of original graph. The edge editing method substantially changes the distance of node properties by linking faraway nodes together or breaking the bridge link of two communities. Clustering method often merges sub graphs to super nodes. Grouping of "similar" nodes are super nodes. The super node represents a "cluster." Then linking between nodes are referred as the edges in between super nodes called "super edges." Each super edge describes more edges in the original graph. Graph based anomaly detection which introduces a model for calculation of regularity graph with rigorous applications to anomaly detection. Graph based anomaly detection refers anomalous substructure detection looks unusual substructures within a entire graph. Secondly anomalous subgraph detection, in which graph is partitioned into set of vertices and it is tested for unusual patterns. Anomalous subgraph and anomalous substructure detection is more important for detecting unusual patterns. In addition conditional entropy measures both regularity and anomaly detection. The novel technique anatomy which releases sensitive data, describes protection of privacy and various correlation of microdata. In addition they introduce linear time algorithm for generating anatomized tables and by decreasing the error reconstruction of the microdata. This innovative technique developed anatomy, preserves privacy and correlation in microdata.

Simple attacks of a K-anonymized dataset have few subtle, but major privacy problems. Firstly they tell an attacker can discover sensitive attributes values. Secondly attackers often have background knowledge. Meanwhile it does not guarantee privacy against hackers. This Framework provides stronger privacy guarantees and describes well represented values for sensitive attributes.

The greedy privacy algorithm introduces that structural information loss measure quantifies loss of information because of edge generalization process. Greedy privacy algorithm can be user balanced that are monitored by other data holders. The Main drawback of this approach denotes that when nodes of similar groups are merged into super node and so relations of various nodes have been lost. Perturbation strategies which refer to Gaussian randomization multiplication and Greedy perturbation algorithm focused on graph theory. The perturbation strategies having same nearest paths and lengths often close to original graph of social network. A perturbation strategy does not perturb edges in which social network structure changes over time. Neighborhood attacks are often having background knowledge of target victim and connection among neighbors. The victim re-identified even victim's sensitive information is preserved by various anonymization techniques. Neighborhood attacks are now in practice. However social networks can answer various aggregate queries. In this paper they are referring only one neighborhood and it is much better to focus on d neighborhoods (d>1) are protected. Eigen values of networks are intimately linked to many topological features.

In this paper, we organized one spectrum randomization approach which automatically improves the graph randomization methods. The utility of graph preserves more privacy protection. Graph spectrum relates many real graphs and provides a perspective edge randomization. A novel approach for anonymizing social network data models offers network structure by allowing samples from that model. This scheme guarantees anonymity by preserving estimation of a wide variety of social network measures. Generalized graph splits the nodes and showed that wide range of graph analysis can be measured accurately and by protecting against risk reidentification. Graph based anonymity notion prevents the identity disclosure of users profiles in which an attacker often have certain prior knowledge. However this method unsuits for social network graphs because in a graph nodes and edges are being

related to each other. A single modification of an edge and node can spread to entire social network. Finally, measuring the utility of a graph becomes more typical. They are not aware of effective metrics of various information loss occurred by the modified nodes and edges.

A new set of various techniques for anonymizing social network data based on merging the entities into classes and by mapping the entities and nodes often denoted them in anonymized target graph. Anonymization techniques are being a challenge to attackers with larger background knowledge information. Yet turned unsatisfied because it has lower utility and less graph structure is here revealed.

Random link attacks (RLAs) performs multiple false identities and creates interactions among various users profiles to attack regular users of social networks. We have showed that RLA attackers can be splitted by their spectral characteristics. Random link attack is a special collaborative attack. The malicious user has complete control by breaking nodes and captures them to attack a more number of randomly chosen victim nodes. Our spectrum detection approach works when hackers choose random victims or by attacking few victims while performing their collaborative attacks. The state of analyzation of privacy protection in social network graphs describes effective anonymization attacks to protect from hackers.

In this paper, starclique, a minimal graph required k-anonymity, where user is identified for all possible contributions of data objects. The identification of social intersection attack can compromise users to identify shared objects relying on social graph topology.

III. PROPOSED WORK

Finally in our proposed approach, we are developing KDLD sequence for target node creation of social network graphs. Given a graph G and its degree sequence consists of triplet namely node position, degree and sensitive labels. KDLD SEQUENCE GENERATION: Given the sensitive degree sequence P and two integers k and L, computes a KDLD sequence. To obtain a new KDLD sequence, same group nodes are needed to be modified for next graph construction process. We further employ two algorithms:

- K-L BASED
- L-K BASED

The algorithms are keeping the nodes of similar degrees to same group to reduce node reidentification process. Algorithm K -L-BASED chooses firstly K elements in original social graph and by monitoring the next element into current group until L-diversity constraint is satisfied.

- Cnew: The cost of developing a new group for the next k elements.
- Cmerge: The cost of merging the next element into the current group.

In this way, target node generation can be created and after that graph construction process is to be generated as follows. Graph construction: Neighborhood_Edge_Editing(): Neighborhood operation describes by adding or by deleting the nodes and edges in the KDLD sequence generation. By doing this modification sensitive labels are being protected from hackers. Adding_Node_Decrease_Degree(): If the node degree is larger than target KDLD sequence generation node, we need to decrease the degree of node by breaking the links between two hop neighbors and by making a direct links to noise nodes. Adding_Node_Increase_Degree(): If the node degree is smaller than target KDLD sequence generation node, we need to increase the degree of node by connecting the links between two hop neighbors and by breaking a direct links to noise nodes. New_Node_Degree_Setting(): This operation describes by assigning degrees to noise nodes. Suppose whose noise node degree is an even number, we select an even degree or if it is odd degree we have to assign odd degree for target nodes. New_Node_Label_Setting(): The final step is to assign labels to newly modified social network graphs. By doing this it is more helpful for preserving distances between labels and remaining labels in social network graphs.

IV GRAPH CONSTRUCTION

In this graph construction, each node and labels details are clearly predicted and by doing adjustments like adding node increase degree, adding node decrease degree, new node label setting and new node degree setting and by doing operations and making use of noise nodes.

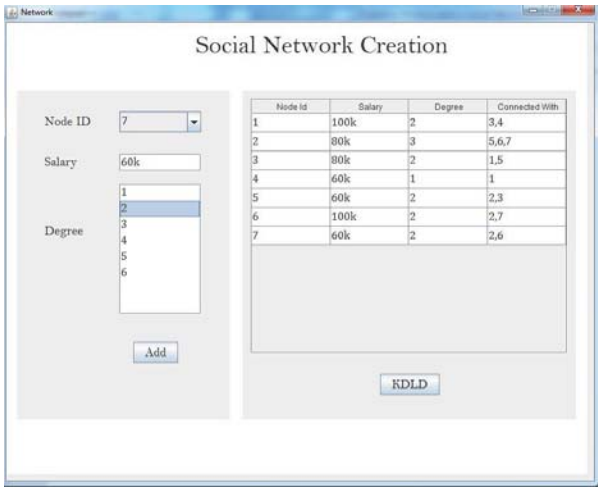

Figure 1. Graph Construction

We propose a k-degree-l-diversity model for privacy preserving social network data publishing. We implement both distinct l-diversity and recursive (c, l)-diversity. In order to achieve the requirement of k-degree-l-diversity, we design 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. We give a rigorous analysis of the theoretical bounds on the number of noise nodes added and their impacts on an important graph property. Our 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.

IV.CONCLUSION

In this paper, we surveyed a few recent studies on anonymization techniques for privacy preserving publishing of social network data.

We propose a k-degree-l-diversity model for privacy preserving social network data publishing. We implement both distinct l-diversity and recursive (*c, l*)-diversity. In order to achieve the requirement of k-degree-l-diversity, we design 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. We give a rigorous analysis of the theoretical bounds on the number of noise nodes added and their impacts on an important graph property. Our 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.

REFERENCES

[1] A. Cam pan and T.M. Truta, "A Clustering Approach for Data and Structural Anonymity in Social Networks," Proc. Second ACM SIGKDD Int'l Workshop Privacy, Security, and Trust in KDD (PinKDD '08), 2008.

[2] E.M. Knorr, R.T. Ng, and V. Tucakov, "Distance-Based Outliers: Algorithms and Applications," The VLDB J., vol. 8, pp. 237-253, Feb. 2000.

[3] 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
[4] 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
[5] L. Liu, J. Wang, J. Liu, and J. Zhang, "Privacy Preserving in Social Networks against Sensi
[6] S. Bhagat, G. Cormode, B. Krishnamurthy, and D. Srivastava,"Class-Based Graph Anonymization for Social Network Data,"Proc. VLDB Endowment, vol. 2, pp. 766-777, 2009.
[7] S. Bhagat, G. Cormode, B. Krishnamurthy, and D. Srivastava,"Class-Based Graph Anonymization for Social Network Data,"Proc. VLDB Endowment, vol. 2, pp. 766-777, 2009.