www.ijecs.in
International Journal Of Engineering And Computer Science ISSN:2319-7242
Volume 3 Issue 12 December 2014, Page No. 9732-9735

Preventing Senstive Social Network Data By Using UINN Algorithm

[1]Tummala Surya Padma [2].B.Venkata Reddy

(M.Tech),
,Asst.Professor.
1,2Dept. of CSE, Kakinada Institute of Technology& Science(KITS), Divili, E.G.Dt, AP,
India

Abstract: Publishing or sharing the social network data for social science research and business analysis lack of privacy. Existing technique k-anonymity is used to prevent identification of microdata. Even though an attacker may gaining sensitive data if a group of nodes largely share the same sensitive labels. We propose an algorithm, universal –match based Indirect Noise Node which makes use of noise nodes to preserve utilities of the original graph. Finally that technique prevents an attacker from reidentifying a user and finding the fact that a certain user has a specific sensitive value.

1. Introduction

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 proposed various privacy models with the corresponding protection mechanisms that prevent both inadvertent private information leakage and attacks by malicious adversaries. These early privacy models are mostly concerned with identity and link disclosure. The social networks are modeled as graphs in which users are nodes and social connections are edges. The threat definitions and protection mechanisms leverage structural properties of the graph. This paper is motivated by the recognition of the need for a finer grain and more personalized privacy.

Users entrust social networks such as Facebook and LinkedIn with a wealth of personal information such as their age, address, current location or political orientation. We refer to these details and messages as features in the user's profile. We propose a privacy protection scheme that not only prevents the disclosure of identity of users but also the disclosure of selected features in users' profiles. An individual user can select which features of her profile she wishes to conceal. 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. The graph representing a small subset of such a social network. Each node in the graph represents a user, and the edge between two nodes represents the fact that the two persons are friends. Labels annotated to the nodes show the locations of users. Each letter represents a city name as a label for each node. Some individuals do not mind their residence being known by the others, but some do, for various reasons. In such case, the privacy of their labels should be protected at data release. Therefore the locations are either sensitive labels or non-sensitive.

The privacy issue arises from the disclosure of sensitive labels. One might suggest that such labels should be simply deleted. Still, such a solution would present an incomplete view of the network and may hide interesting statistical information that does not threaten privacy. A more sophisticated approach consists in releasing information about sensitive labels, while ensuring that the identities of users are protected from privacy threats. We consider such threats as neighborhood attack, in which an adversary finds out sensitive information based on prior knowledge of the number of neighbors of a target node and the labels of these neighbors.

2. Related Work

The first necessary anonymization technique in both the contexts of micro- and network data consists in removing identification. This nave technique has quickly been recognized as failing to protect privacy. For microdata, Sweeney et al. propose k-anonymity to circumvent possible identity disclosure in naively anonymized microdata. L-diversity is proposed in in order to further prevent attribute disclosure.

Similarly for network data, Backstrom et al., in [2], show that naive anonymization is insufficient as the structure of the released graph may reveal the identity of the individuals corresponding to the nodes. Hay et al. [3] emphasize this problem and quantify the risk of re identification by adversaries with external information that is formalized into structural queries Recognizing the problem, several works [5, 11, 18, 20{4,5] propose techniques that can be applied to the naive anonymized graph, further modifying the graph in

order to provide certain privacy guarantee. Some works are based on graph models other than simple graph [6,7].

To our knowledge, Zhou and Pei [8,9] and Yuan et al. [10] were the first to consider modeling social networks as labeled graphs, similarly to what we con-sider in this paper. To prevent reidentication attacks by adversaries with immediate neighborhood structural knowledge, Zhou and Pei [11] propose a method that groups nodes and anonymizes the neighborhoods of nodes in the same group by generalizing node labels and adding edges. They enforce a k-anonymity privacy constraint on the graph, each node of which is guaranteed to have the same immediate neighborhood structure with other k-1 nodes. In [12], they improve the privacy guarantee provided by k-anonymity with the idea of `l-diversity, to protect labels on nodes as well. Yuan et al. [13] try to be more practical by considering users' different privacy concerns. They divide privacy requirements into three levels, and suggest methods to generalize labels and modify structure corresponding to every privacy demand. Nevertheless, neither Zhou and Pei, nor Yuan et al. consider labels as a part of the background knowledge. However, in case adversaries hold label information, the methods of cannot achieve the same privacy guarantee. Moreover, as with the context of microdata, a graph that satisfies a k-anonymity privacy guarantee may still leak sensitive information regarding its labels.

LITERATURE SURVEY:

A network data set is a graph representing a set of entities and the connections between them. Network data can describe a variety of domains: a social network describes individuals connected by personal relationships; an information network might describe a set of articles connected by citations; a communication network might describe Internet hosts related by traffic flows. As our ability to collect network data has increased, so too has the importance of analyzing these networks. Networks are analyzed in many ways: to study disease transmission, to measure the influence of a publication, and to evaluate the network's resiliency to faults and attacks. Such analyses inform our understanding of network structure and function.

However, many networks contain highly sensitive data. For example, Potterat et al. published a social network which shows a set of individuals related by sexual contacts and shared drug injections. While society knows more about how HIV spreads because this network was published and analyzed, researchers had to weigh that benefit against possible losses of privacy to the individuals involved Without clear knowledge of potential attacks. Other kinds of networks, such as communication networks are also considered sensitive. The sensitivity of the data often prevents the data owner from publishing it. For example, to our knowledge, the sole publicly available network of email communication was published only because of government litigation.

The objective of the data owner is to publish the data in such a way that permits useful analysis yet avoids disclosing sensitive information. Because network analysis can be performed in the absence of entity identifiers, the data owner first replaces identifying attributes with synthetic identifiers. We refer to this procedure as naive anonymization. It is a common practice in many domains, and it is often implemented by simply encrypting identifiers. Presumably, it protects sensitive information because it breaks the association between the sensitive data and real-world individuals.Social networks have been studied for a century [Sim08] and are a staple of research in disciplines such as epidemiology, sociology, economics and many others. The recent proliferation of online social networks such as MySpace, Facebook, Twitter. Even in the few online networks that are completely open, there is a disconnect between users' willingness to share information and their reaction to unintended parties viewing or using this information. Most operators thus provide at least some privacy controls. Many online and virtually all offline networks (*e.g.*, telephone calls, email and instant messages, *etc.*) restrict access to the information about individual members and their relationships. Network owners often share this information with advertising partners and other third parties. Such sharing is the foundation of the business case for many online socialnetwork operators.

However, naive anonymization may be insufficient. A distinctive threat in network data is that an entity's connections (i.e., the network structure around it) can be distinguishing, and may be used to re-identify an otherwise anonymous individual. we investigate the threat of structural re-identification in anonymized networks. We consider how a malicious individual (the adversary) might learn about the network structure and then attempt to reidentify entities in the anonymized network. We formally model adversary capabilities, demonstrate successful attacks on real networks, and propose an improved anonymization technique First, we survey the current state of data sharing in social networks, the intended purpose of each type of sharing, the resulting privacy risks, and the wide availability of auxiliary information which can aid the attacker in de-anonymization.

Second, we formally define privacy in social networks and relate it to node anonymity.We identify several categories of attacks, differentiated by attackers' resources and auxiliary information. We also give a methodology for measuring the extent of privacy breaches in social networks, which is an interesting problem in its own right.

Third, we develop a generic re-identification algorithm for anonymized social networks.

Fourth, we give a concrete demonstration of how our deanonymization algorithm works by applying it to Flickr and Twitter, two large, real-world online social networks.

3. Existing System:

Recently, much work has been done on anonymizing tabular microdata. A variety of privacy models as well as anonymization algorithms have been developed (e.g., kanonymity, l-diversity, t-closeness. 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.

DISADVANTAGES OF EXISTING SYSTEM:

- 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.
- 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.

PROPOSED SYSTEM:

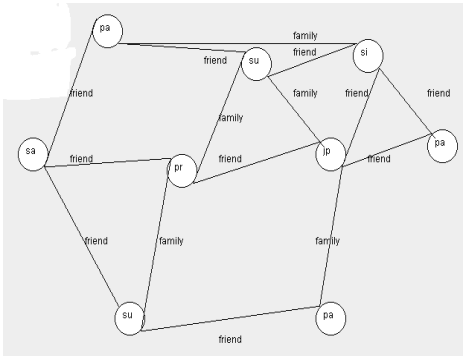
We propose a novel idea to preserve important graph properties, such as distances between nodes by adding certain "noise" nodes into a graph. This idea is based on the following key observation.

In Our proposed system, 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, we define a Universal Match-based indirect noise node 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.

ADVANTAGES OF PROPOSED SYSTEM:

- ❖ We use Universal Match-based indirect noise node to prevent not only the reidentification of individual nodes but also the revelation of a sensitive attribute associated with each node.
- ❖ We propose a novel graph construction technique which makes use of noise nodes to preserve utilities of the original graph. Two key properties are considered: 1) Add as few noise edges as possible; 2) Change the distance between nodes as less as possible.
- ❖ We present analytical results to show the relationship between the number of noise nodes added and their impacts on an important graph property.

4. System Architecture



Social Network Users Representation As Graph

Algorithm

The main objective of the algorithms that we propose is to make suitable grouping of nodes, and appropriate modification of neighbors' labels of nodes of each group to satisfy the *l-sensitive-label-diversity* requirement. We want to group nodes with as similar neighborhood information as possible so that we can change as few labels as possible and add as few noisy nodes as possible. We propose an algorithm, Universal Match-based Indirect Noise Node that does not attempt to heuristically prune the similarity computation as the other two algorithms, Direct Noisy Node Algorithm (DNN) and Indirect Noisy Node Algorithm (INN) do. Algorithm *DNN* and *INN*, which we devise first, sort nodes by degree and compare neighborhood information of nodes with similar degree. Details about algorithm *DNN* and *INN*.

Algorithm UINN

The algorithm starts out with group formation, during which all nodes that have not yet been grouped are taken into consideration, in clustering-like fashion. In the first run, two nodes with the maximum similarity of their neighborhood labels are grouped together. Their neighbor labels are modified to be the same immediately so that nodes in one group always have the same neighbor labels. For two nodes, v_1 with neighborhood label set (LS_{v1}), and v_2 with neighborhood label set (LS_{v2}), we calculate neighborhood label similarity (NLS) as follows:

$$\text{NLS}(v_1, v_2) = \frac{|LS_{v_1} \cap LS_{v_2}|}{|LS_{v_1} \cup LS_{v_2}|}$$

Larger value indicates larger similarity of the two neighborhoods.

Then nodes having the maximum similarity with any node in the group are clustered into the group till the group has ` nodes with different sensitive labels. Thereafter, the algorithm proceeds to create the next group. If fewer than ` nodes are left after the last group's formation, these remainder nodes are clustered into existing groups according to the similarities between nodes and groups.

After having formed these groups, we need to ensure that each group's members are indistinguishable in terms of *neighborhood information*. Thus, neighborhood labels are modified after every grouping operation, so that labels of nodes can be accordingly updated immediately for the next grouping operation. This modification process ensures that all nodes in a group have the same *neighborhood information*. The objective is achieved by a series of modification operations. To modify graph with as low information loss as possible, we devise three modification operations: *label union*, *edge insertion* and *noise node addition*. Label union and edge insertion among nearby nodes are preferred to node addition, as they incur less alteration to the overall graph structure.

Edge insertion is to complement for both a missing label and insufficient degree value. A node is linked to an existing nearby (two-hop away) node with that label. Label union adds the missing label values by creating super-values shared among labels of nodes. The labels of two or more nodes coalesce their values to a single super-label value, being the union of their values. This approach maintains data integrity, in the sense that the true label of node is included among the values of its label super-value. After such edge insertion and label union operations, if there are nodes in a group still having different neighborhood information, noise nodes with non-sensitive labels are added into the graph so as to render the nodes in group indistinguishable in terms of their neighbors' labels. We consider the unification of two nodes' neighborhood labels as an example. One node may need a noisy node to be added as its immediate neighbor since it does not have a neighbor with certain label that the other node has; such a label on the other node may not be modifiable, as its is already connected to another sensitive node, which prevents the re-modification on existing modified groups.

5. Modules

MODULES DESCRIPTION:

User Pane

User Pane is a block of information about a given user, like those typically found on a forum post, but can be used in other places as well. From core, it collects the user picture, name, join date, online status, contact link, and profile information. In addition, any module or theme can feed it more information via the preprocess system. All of this information is then gathered and displayed using a template file.

User Relationships

In this module we develop User Relationship module. Allows users to create named relationships between each other. It is the basic building block for a social networking site, or any site where users are aware of one another, and communicate.

Content Access

This module allows you to manage permissions for content types by role and author. It allows you to specify custom view, edit and delete permissions for each content type. Optionally you can enable per content access settings, so you can customize the access for each content node.

Protection of structural information

Our privacy preserving goal is to prevent an attacker from re-identifying a user and finding the fact that a certain user has a specific sensitive value.

6. Conclusion

In this paper, we propose Universal Match –Based Technique for privacy preserving social network data publishing. 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.

References

[1] L. Backstrom, C. Dwork, and J.M. Kleinberg, "Wherefore Art Thou r3579x?: Anonymized Social Networks, Hidden Patterns, and Structural Steganography," Proc. Int'l Conf. World Wide Web (WWW), pp. 181-190, 2007.

[2] 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.

[3] A. Campan 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.

[4] 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.

[5] 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.

[6] S. Das, O. Egecioglu, and A.E. Abbadi, "Privacy Preserving in Weighted Social Network," Proc. Int'l Conf. Data Eng. (ICDE '10), pp. 904-907, 2010.

[7] 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.

[8] 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.

[9] N. Shrivastava, A. Majumder, and R. Rastogi, "Mining (Social) Network Graphs to Detect Random Link Attacks," Proc. IEEE 24th Int'l Conf. Data Eng. (ICDE '08), pp. 486-495, 2008.

[10] L. Sweeney, "K-Anonymity: A Model for Protecting Privacy," Int'l J. Uncertain. Fuzziness Knowledge-Based Systems, vol. 10, pp. 557- 570, 2002.

[11] X. Ying, X. Wu, and D. Barbara, "Spectrum Based Fraud Detection in Social Networks," Proc. IEEE 27th Int'l Conf. Very Large Databases (VLDB '11), 2011.

[12] X. Ying and X. Wu, "Randomizing Social Networks: A Spectrum Preserving Approach," Proc. Eighth SIAM Conf. Data Mining (SDM '08), 2008.

[13] 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