ISSN:0975-9646

Anonymization: Enhancing Privacy and Security of Sensitive Data of Online Social Networks

Mr.Gaurav .P.R.
PG Student, Dept.Of CS&E
S.J.M.I.T
Chitradurga, India

Mr.Gururaj.T M.Tech
Associate Professor, Dept.Of CS&E
S.J.M.I.T
Chitradurga, India

***Abstract*—Privacy is one of the major concerns when publishing or sharing social network data for social science research and business analysis. This paper is motivated by the recognition of the need for a finer grain and personalized privacy in data publication of social netwoks.Recently , researchers have proposed a privacy protection scheme that not only prevents the disclosure of identity of users but also the disclosure of selected features in user' s profiles. An individual user can select which features of his profiles he 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. We treat node labels both as background knowledge an adversary may possess, and as sensitive information that has to be protected. Privacy protection algorithms that allow for graph data to be published in a form such that an adversary who possesses information about a nodes neighborhood cannot safely infer its identity and its sensitive labels. To this aim, the algorithms transform the original graph into a graph in which nodes are sufficiently indistinguishable. The algorithms are designed to do so while losing as little information and while preserving as much utility as possible. We evaluate empirically the extent to which algorithms preserve the original graph' s structure and properties. We show that our solution is effective, efficient and scalable while offering stronger privacy guarantees than those in previous research. Privacy models evolved w.r.t k-anonymity to prevent node reidentification through structure information.**

Index Terms*—Social networks, data mining, privacy, anonymous. *(Key words)

I. INTRODUCTION

Data mining is the process of automatically discovering useful information in large data repositories. Data mining techniques are deployed to scour large databases in order to find novel and useful patterns that might otherwise remain unknown.

Data mining have encountered traditional data analysis techniques in meeting the challenges posed by new data sets. The following are some of the challenges that motivated the development of data mining; Scalability, High dimensionality, Heterogeneous and Complex Data, Data Ownership and Distribution, Non-traditional Analysis.

Data mining tasks are generally divided into two major categories: Predictive tasks and Descriptive tasks.

Social Network

A social network describes entities and connections between them. The entities are often individuals; they are connected by personal relationships, interactions or flows of information. Social network analysis is concerned with uncovering patterns in the connection between entities. It has been widely applied to organizational networks to classify the influence or popularity of individuals and to detect collusion and fraud. Social network analysis can also be applied to study disease transmission in communities, the functioning of computer networks and emergent behavior of physical and biological systems.

Here in this paper we are concentrating mainly on providing security for sensitive labels of large data sets and repositories. Due to day to day updating and the fast growing worlds of social networks like Facebook, LinkedIn, more researchers are involving in the process of obtaining the information from these social networking data, such as the user behavior, community growth, disease spreading etc.,however at the same time that published social network data should not disclose the private information of individuals. Thus, by protecting individual's privacy and at the same time preserve the utility of social network data becomes a challenging aspect. Here in this paper we are going to explain it through graph models, where each vertex in the graph is associated with sensitive labels.

Recently, much work has been carried out on anonymizing tabular data. A variety of privacy models as well as Anonymization algorithms have been developed like k-anonymity, l-diversity, and t-closeness. In publishing the micro data, some of the non sensitive 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 the 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 sub graph of a node, to identify a node. To prevent structure attacks, a published graph should satisfy the k-anonymity [8],[7]. The goal is to publish social graph, which always has at least k candidates in different attack scenarios in order to protect privacy.

Current approaches for protecting graph privacy can be classified into two categories: clustering[7],[8],[4] and edge editing [1],[2],[8]. **Clustering** is to merge a sub graph to one super node, which is unsuitable for sensitive labeled graphs, since when a group of nodes is 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. Edge-editing may largely destroy the properties of a graph. The edge-editing method

sometimes may change the distance properties substantially by connecting two far away nodes together or deleting the bridge link between two communities.

To summarize, we made the following contributions:

- Combine 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.
- Using of distinct *l*-diversity to demonstrate our algorithm and give the detailed discussion about how more complex recursive (c,*l*)-diversity can be implemented.
- A novel based graph construction technique is proposed 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.

Social networks, patient networks and email networks are all examples of graphs that can be studied to learn about information diffusion community structure and different system processes however; they are also all examples of graphs containing potentially sensitive information. While several Anonymization techniques have been proposed for social network data publishing, they all apply the Anonymization procedure on the entire graph. This project proposes a local Anonymization algorithm that focuses on obscuring structurally important nodes that are not well anonymized. Doing so reduces the cost of the overall anonymization procedure. The current technique reduces the cost of anonymization by an order of magnitude while maintaining and even improving the accuracy of different graph centrality measures, example degree and between's when compared to another well known data publishing approach. This paper then explores the underlying anonymity inherent in the topological structure of online social networks to better understand which parts are not well anonymized. The paper also states that while some components are well anonymized, sub graphs of weak nodes do exist. The paper also finds the measure that better captures the potential impact of nodes in the network being identified by an attacker.

II. PROBLEM DESCRIPTION

In this paper, a social network graph is defined as:

Social Network Graph: a social network 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. E→ V×V is the set of edges between vertices, σ is a set of the labels that vertices have, λ: V→σ maps vertices to their labels.

Here we use the words "node" and "vertex" interchangeably. In a published (privacy preserving) social network graph, an attacker could reidentify a node by degree information and further infer sensitive labels. To prevent this possible leakage, we define "*k*-degree-*l*-diversity" principle for published graphs, which have the same spirit of *k*-*l* diversity in relational data.

KDLD- For each vertex in agraph, there exists at least *k*-1 other vertices having the same degree in the graph. Moreover, the vertices with the same degree contain at least *l* distinct sensitive labels.

A KDLD graph protects two aspects of each user when an attacker uses degree information to attack: 1) The probability that an attacker can correctly reidentify this user is at most 1/k; 2) The sensitive label of this user can at least be related with l different values. Since each equivalent class contains at least k nodes, when an attacker uses the degree to reidentify a node, the probability he correctly reidentifies this user is at most 1/k.

Social networks, patient networks and email networks are all examples of graphs that can be studied to learn about information diffusion, community structure and different system processes; however they are also all examples of graphs containing potentially sensitive information. While there are several anonymization techniques for data publishing, but they all get applied for the entire graph. This paper is supposed to be proposed only on the sensitive labels anonymization obscuring whether structurally important nodes are not well anonymized or not. By doing so reduces the cost of overall anonymization procedure.

III. LITERATURE SURVEY

A.EXISTING SYSTEM

The current trend in the Social Network it not giving the privacy about user profile views. The method of data sharing or (Posting) has taking more time and not under the certain condition of displaying sensitive and non sensitive data. Initially k-degree anonymity was enough to prevent sensitive labels from structure attacks However when time passed and in many applications of social network where each node got its own sensitive attributes which was supposed to be published. For example, a graph may contain the user salaries which are sensitive. In this case, k-degree alone cannot prevent the inference of sensitive attributes of individuals. Therefore, when sensitive labels are considered, the l-diversity should be adopted for graphs. Current approaches for protecting graph privacy are classified into two categories:

- Clustering
- Edge editing

Clustering is to merge a sub graph to one super node, which is unsuitable for sensitive labeled graphs, since when a group of nodes is 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.

Data refers to organized personal information in the form of rows and columns. Row refers to individual tuple or record and column refers to the field. Tuple that forms a part of a single table are not necessarily unique. Column of a table is referred to as attribute that refers to the field of information, thereby an attribute can be concluded as domain. It is necessary that attribute that forms a part of the table should be unique. According to L.Sweeney et.al.,(2002)[2] each row in a table is an ordered n-tuple of values $<d_1,d_2,....d_n>$ such that each value dj forms a part of the domain of j^{th} column for j=1,2,.....n where 'n' denotes the number of columns.

ATTRIBUTES

Consider a relation R(a1,a2,...an) with finite set of tuples.Then the finite set of attributes of R are {a1,a2,a3...,an},provided a table R(a1,a2,...an),{a1,a2,...an} {a1,a2,...an} and a tuple 1 R,1[ai,...an] corresponds to ordered set of values vi,...vj of ai...aj in 1.R[ai...aj] corresponds to projection of attribute values a1,a2,...an in R,thereby maintaining tuple duplicates. According to Ningui Li, Tiancheng Li et.al,[3](2010),attributes among itself can be divided into 3 categories namely

1. Explicit identifiers-Attributes that clearly identifies individuals. For e.g., Social Security Number for US citizen.

2. Quasi identifiers-Attributes whose values when taken together can potentially identify an individual.Eg, postal code, age, sex of a person. Combination of these can lead to disclosure of personal information.

3. Sensitive identifiers-That are attributes needed to be supplied for researchers keeping the identifiers anonymous. For e.g., 'disease' attribute in a hospital database, 'salary' attribute in an employee database.

QUASI-IDENTIFIERS

As proposed by L.Sweeney et.al., (2001)[2], A single attribute or a set of attributes that, in combination with some outside world information that can identify a single individual tuple in a relation is termed as quasi-identifier. Given a set of entities E, and a table B(a1,....,an0, f_a:E→B and f_b:B→E'.A quasi-identifier of B,written as UE, is a set of attributes $\{a_i,...,a_j\}\rightarrow\{a_1,...,a_n\}$ where: s_i U such that $f_a(f_b(s_i)[U_E])=s_i$.

k-ANONYMITY

Let RT (A1,A2,...,An) be a table and QI_{RT} be the Quasi identifier. RT is said to be k-anonymous[2] if and only if each sequence of values in $RT[QI_{RT}]$ appears at least k-times in $RT[QI_{RT}]$.In short, the quasi identifier must appear at least 'k'times in RT, where k=1,2,3,... where 'k' is termed to be the anonymity of the table.

***l*-DIVERSITY**

Since k-Anonymity failed to secure the attribute disclosure, and is susceptible to homogeneity attack and background knowledge attack introduced a new privacy notation called '*l*-diversity' [2].

Problems in existing system:

1. There is no way to publish the non sensitive data to all in social network.
2. It is not providing privacy about user's profiles.
3. Some mechanisms prevent both inadvertent private information leakage and attacks by malicious adversaries.
4. Edge-editing may largely destroy the properties of the graph and it may sometimes also change the distance properties.
5. Mining over these sensitive data might get wrong conclusions.
6. Solely relying on the edge editing may not be the good solution to preserve data utility.

B.PROPOSED SYSTEM

Here, we extend the existing definitions of modules and we introduced the sensitive or non-sensitive label concept in our project. We overcome the existing system disadvantages in our project. To address the issues of existing 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. 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. This paper proposes 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.

Advantages of the proposed system:

1. Here we can publish and post both sensitive and non sensitive data to everyone in social network like ads or jobs.
2. Privacy is provided for user's profile in such a way that unwanted persons are not able to view profiles.
3. Adding noise nodes is carried out but the distance between the original nodes are mostly preserved.
4. Privacy preserving goal is that to prevent an attacker from reidentifying a user and finding the fact that a certain user has a specific sensitive value.
5. To achieve this goal, we 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.
6. Low overhead.

C.EQUATIONS

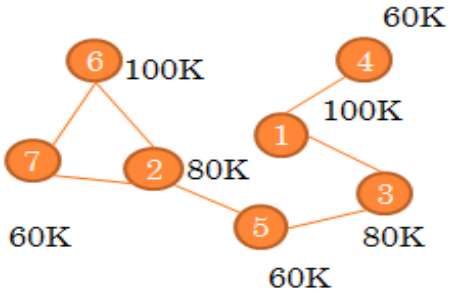
The social distance between all node pairs of a graph is measured by average shortest path length (APL). APL is a concept in network topology that is defined as the average of distances between all pairs of nodes. It is measure of the efficiency of information or mass transport on a network. Some queries like "the nearest node for a group of nodes" are related to APL.

The APL of a graph G is

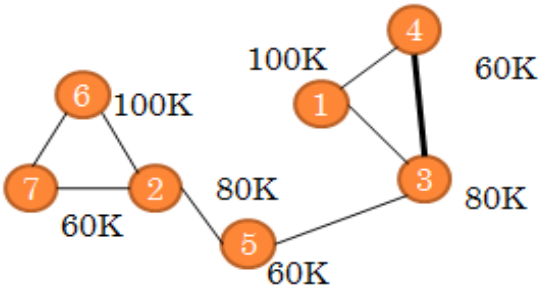
$$APL_G = \frac{2}{N(N-1)}\sum_{\forall n_i, n_j \in G} d(n_i, n_j),$$

Where in, d (n_i, n_j) is the length of the shortest path between nodes n_i and n_j, N is the number of nodes in the graph.

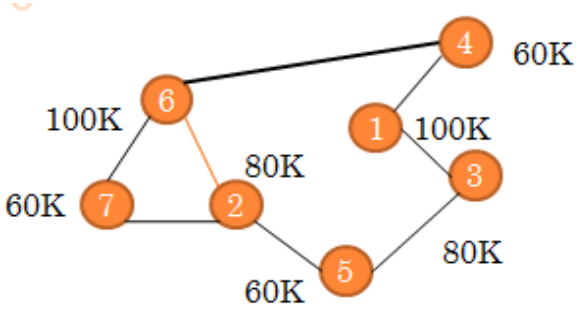
IV. FIGURES



(a)Original graph

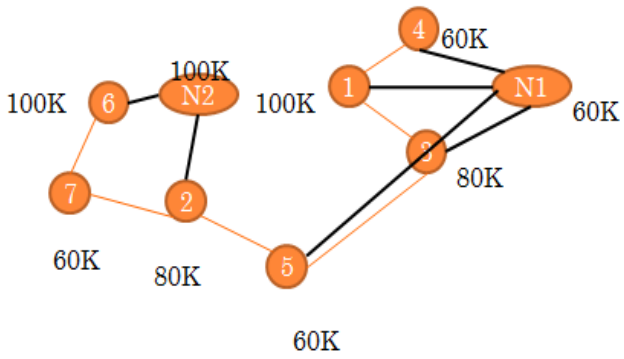


(b) 2-degree anonymous graph



(c) 2-degree-2-diversity graph

Fig. 1. Publish a graph with degree and label anonymity



(d) 2-degree-2-diversity by adding noise node

Fig. 2. Example for adding noise node.

IV. GRAPH CONSTRUCTION

Algorithm Skeleton

The algorithm consists of five steps:

- Step 1:Neighborhood Edge Editing()
  We add or delete some edges if the corresponding edge-editing operation follows the neighborhood rule. By doing this, the sensitive degree sequence of original graph is preserved;
- Step 2:Adding Node Decrease Degree()
  For any node whose degree is larger than its target degree, we increase its degree to the target degree by making using of noise nodes;
- Step 3:Adding Node Increase Degree()
  For any node whose degree is smaller than its target degree, 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, we do some adjustment to make it has a degree. Then, the noise nodes are added into the same degree groups;
- Step 5:New Node Label Setting()
  We assign 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 of the published graph is a KDLD sequence.

IMPLEMENTATION DETAILS

1. Data Collection
2. Attribute Selection
3. Tuple Sorting
4. K-Anonymity

A. DATA COLLECTION

Data collection module collect the data from hospital repository, data consist of age, gender, zip, income marital status and disease , A dataset contains a number of rows where each row is represented by a tuple T. The dataset is made up of several attributes, which are composed of identifiers, quasi-identifiers and non-identifiers. Our main concern here is the attributes of quasi-identifiers.

B. ATTRIBUTE SELECTION

Sensitive attributes are sorted in descent order according to amount of tuple owning highly sensitive value of each attribute. For example, N1 tuple own high sensitive values of SAi, N2 tuple own high sensitive values of SAj. if N1> N2, SAi stands before SAj. We renumber sensitive attributes in order. For arbitrary sensitive attributes SAi and SAj, i<j means Ni>Nj.

C. TUPLE SORTING

Tuple Sorting sort tuple in dataset based on greedy strategy. Firstly entire dataset is divided into two groups. Tuple own highly sensitive values of SA1 are put into group 1. Group 2 contains others tuple. Similarly two groups are divided according to SA2 respectively. Tuple sorting divide groups recursively until SAM has been checked. Detail of this step is shown in function *sorting*. *n* is the amount of tuple to be

sorted. ti(I) is the *I*th sensitive value of tuple ti. Symbol '+' joins two lists into one.

D. K- anonymity

A module which implements the anonymization of Optimal K-anonymity algorithm.

Along with k-anonymity we are going to implement the KDLD which is still more efficient than that of the k-anonymity and will overcome the disadvantages of the previous algorithms and methods which are used to provide privacy for the sensitive labels in a social networks and also we make use of recursive (*c*,*l*)-diversity to assign sensitive labels to noise nodes.

V. CITING PREVIOUS WORK

We make detailed investigations on a spectrum of privacy models and graphical model where the node of a graph indicates a sensitive attribute. Recently a lot of works have been done on anonymizing a relational database.

k-anonymity approach developed by L.Sweeney et.al.,(2002)[2], a model for protecting privacy which poses the condition that a database to be k-anonymous, then each record is indistinguishable from at least k-l other records with respect to their quasi-identifiers.Quasi-Identifiers are attributes whose values when taken together can potentially identify an individual. Since k-anonymity failed to secure the attribute disclosure, and is susceptible to homogeneity attack and background knowledge attack A.Machanavajjhala et.al.,[5](2007) introduced a new privacy notation called l-diversity. An equivalence class is said to possess l-diversity if there are at least 'l' well represented values for the sensitive attribute. A table is said to have l-diversity if every equivalence class of the table has l-diversity. Privacy is measured by the information gain of an observer. Before seeing the released table the observer may think that something might happen to sensitive attribute value of a single person. After seeing the released table the observer may have the details about the sensitive attributes. T-closeness should have the distance between the class and the whole table is no more than a threshold t, Ningui Li et.al.,[3](2010).

Graph structures are also published hand-in-hand when publishing social network data as it may be exploited to compromise privacy. The degree and sub graph of a node could be used to identify a node. It is observed from literature that in order to prevent structure attacks the graph is enforced to satisfy k-anonymity.

In the previous base papers only the concepts regarding k-anonymity was discussed and implemented to protect sensitive labels from adversaries, but the model failed in providing privacy n security for sensitive labels henceforth in the papers [2] & [3] the concept of k-anonymity and *l*-diversity is discussed and stated. Finally in this paper we have made use of KDLD sequence keeping few journals as references [4], [5], [6].

Simply removing identifiers in social networks does not guarantee privacy. The unique patterns, such as node degree or sub graph to special nodes, can be used to reidentify the nodes. The attack that uses certain background knowledge to reidentify the nodes/links in the published graph called "passive attack". There are two models proposed here to publish privacy preserved graph: edge-editing based model and clustering –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. 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 the super nodes which is called "super edges." Each super edge may represent more than one edge in the original graph. We call the graph that only contains super nodes and super edges as a clustered graph. Most edge-editing-based graph protection models implement k-anonymity of nodes on different background knowledge of the attacker. Liu and Terzi [7] 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 [8] considered k-neighborhood anonymous model: for every node, there exist at least other k-1 nodes sharing isomorphic neighborhoods. In paper, the k- neighborhood anonymity model is extended to k-neighborhood-l-diversity model to protect the sensitive node label. A graph is k-Automorphism if and only if for every node there exist at least k-1 other nodes that do not have any structure with it.

Besides the "passive attack," there's another type of attack on social networks, which is called "active attack". Active attack when tack is to actively embed special sub graphs into a social network when this social network is collecting data. An attacker can attack the users who are connected with the embedded sub graphs by reidentifying these special sub graphs in the published graph. One method to prevent active attack is to recognize the fake nodes added by the attackers and remove them before publishing the data. To identify fake nodes, triangle probability difference between normal nodes and fake nodes is found out and also spectrum analysis method is used.

VI. DISCUSSION

For stronger graph protection models such as k-neighborhood anonymity, it is also helpful to preserve Average shortest Path Length by carefully adding some nodes. A graph is k-neighborhood anonymous if: for every node there exist at least other k-1 nodes sharing isomorphic neighborhood graph.

The basic procedure to generate a k-neighborhood anonymous [8] graph is: 1) Sort all nodes by their neighborhood graph size in descending order;2) Recursively adjust two nodes neighborhood graphs to be the same until k-anonymity graph is generated. When adjusting neighborhood graphs G_u and G_v with $|G_u|>|G_v|$ to be the same, new nodes should be introduced into G_v. An unanonymized node with the smallest degree has the highest priority to be added. The noise node adding strategy should be considered in this step to improve the utility of the published graph.

VII. CONCLUSION

In this paper, 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 and a rigorous analysis of the theoretical bounds are given on the number of the noise nodes added and their impacts on an important graph property. Protocols should be designed to help these publishers publish a unified data together to guarantee the privacy.

REFERENCES

[1] Protecting Sensitive Labels in social Network Data Anonymization by Mingxuan Yuan, Lei Chen, *Member*, IEEE, Philip S.Yu, *Fellow, IEEE, and* Ting Yu-March 2013

[2] L.Sweeney,"K-Anonymity: A Model for Protecting Privacy", Int'l J.uncertain. Fuzziness Knowledge-Based Systems, 2002

[3] N.Li and T.Li,"T-closeness: Privacy beyond K-Anonymity and L-Diversity"IEEE 23rd Int'l Conf.Data Eng, 2010

[4] International Journal of Innovative Research in Computer and Communication engineering, by S.charanyaa and Prof. T. Shanmugapriya October 2013.

[5] A.Machanavajjhala et.al.,"l-diversity:Privacy beyond K-anonymity,ACM Trans Knowledge Discovery Data(2007)

[6] Anonymizing Social Networks Michael Hay, Gerome Miklau, David Jensen, Philipp Weismann Siddarth Srivastava, University Of Massachusetts Amherst Computer Science Department, and March 2007.

[7] K. Liu and E. Terzi,"Towards Identity Anonymization on Graphs", *SIGMOD'08:Proc.ACM SIGMOD Int'l conf.Management of Data*, pp93-106, 2008

[8] 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