*International Journal of scientific research and management (IJSRM)*
*\||Volume||3||Issue||7||Pages|| 3252-3258||2015||*
*Website: www.ijsrm.in ISSN (e): 2321-3418*

# Privacy Preservation of Social Network Data

***Miss. Srushti. N. Patil, Prof. Priti. A. Khodke***
Master of Engineering 2nd yr.
P.G. Department of Computer Science and Engineering
Prof. Ram Meghe College of Engineering & Management
Badnera, India
srushpatil@gmail.com
Head Of Department
P.G. Department of Computer Science and Engineering
Prof. Ram Meghe College of Engineering & Management
Badnera, India
priti.khodke@gmail.com

Abstract: The publication of social network data necessitates a privacy threat for their users. The emerging popularity ofsocial networks has generated fascinating data management and data mining problems. Such social networks are of interest toresearchers from several disciplines like sociology, psychology, marketing research, or epidemiology. Now a day, a largeamount of social network data has been released in different ways. Since social networks usually contain private information,a vital concern in publishing these data for study is their privacy. Sensitive information related to users of social networksshould be protected. Having some local knowledge concerning users of a social network, an attacker may easily thrash theprivacy of some victims. Simply removing the identities of nodes before releasing the social network data does not guaranteeprivacy. Hence, it is required to anonymize the data before its publication so as to address the need to respect the privacy ofthe individuals whose sensitive information is included in the data. Typically Data anonymization trades off with utility.Hence, it is needed to find a golden path in which the released anonymized data still holds enough utility, while protecting privacy to some accepted degree. So, the summons is to implement methods to release social network data in a way that affords utility without compromising privacy.

Keywords: Social Network, Privacy, Sensitive Information, Data Anonymization.

## 1. Introduction

Privacy is the ability of a user or a group to concealed themselves, or information regarding themselves, and thereby express themselves selectively. When something is private to the user, this implies that there is something special or sensitive to them. The boundaries and content of what is examined private conflict among individuals, however share common themes. [1].

An online social network is outlined as a web-based service that enables users to "construct a public or semi-public profile within a finite system; articulate a listing of other users with whom they split a connection; and sight and traverse their list of connections and those created by others in the system"[2]. A social network is a collection of social entities (such as individuals or organizations) and relationships between them. These entities, or nodes, are abstract representations of entities that are connected by one or more attributes. The connections or edges, corresponds to relationships between these nodes [3].

Users entrust social networks like Facebook and LinkedIn with a wealth of personal information such as their age, address, current location etc. These are called as features in the user's profile [4]. Basically, social networks

are modelled as a graph, where the nodes of the graph represent the entities and edges represent relationship between them. Besides entities and edges, extra information regarding users and relationships can be represented by labels. Vertex label correspond to features in user's profile. Edge label correspond to weight of a relationship that defines a quantitative measure of the relationship e.g. the degree of a friendship etc. and type of a relationship that states the nature of the relationship, e.g. friendship, an email, web links [5].

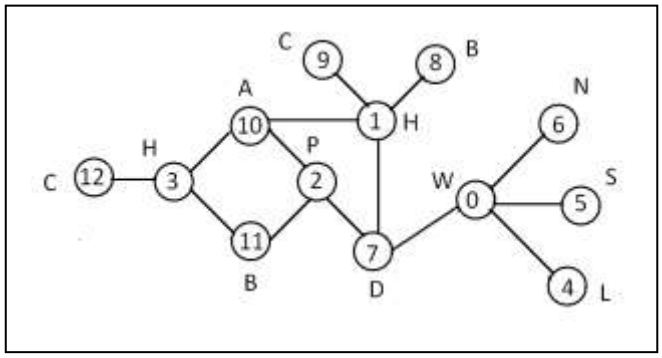


**Figure 1:** Example of the labeled graph representing a social network

Social networking sites have a large number of etc registered individuals, and each individual is associated with a variety of others through friendship, professional association (being members of communities), and so on [6]. The information published by an individual on a social network can be viewed by all his friends on a social network. A large amount of data is being published on social networks there is a necessity to protect the privacy of a network.

Due to the popularity of social networks, several approaches have been proposed to secure the privacy of the social networks. The prior works suppose that the attacks exploit the same background knowledge. But, in practice, different users have different privacy protection requirements. Thus, assuming the attacks with the identical background knowledge do not meet the personalized privacy protection demands, simultaneously, it loses the possibility to attain better utility by taking benefit of differences of user's privacy preservation demands [7].

## 2. Existing Strategies For Privacy Protection On Social Networks

This section discusses six existing strategies for privacy protection on social networks along with proposed method, Description of method, Data to be protected, on which social network it is applicable, Data-set which is used, Advantages and Disadvantages are discussed in this section.

### 2.1. Sensitive Label Privacy Protection on Social Network Data [4]:

Song, Panagiotis Kara's, Qian Xiao, and StéphaneBressan [4] considered node labels both as background knowledge an adversary may have as well as sensitive information that needs to be secured. They proposed privacy protection algorithmsthat permit for graph data to be released in a way such that an attacker who has information regarding a node's neighborhood cannot safely conclude its identity and its sensitive labels. Two nodes are expected to have the same labels of neighbors and are within two hops (having common neighbors), a single node is added. In other words, they merge some noisy nodes with the same label, thus resulting in fewer noisy nodes. To the present aim, the algorithm converts the original graph into one in which vertices are sufficiently indistinguishable. The algorithms are designed to do so while losing as small information and while protecting as much utility as possible. They evaluate empirically the extent to which the algorithms protect the original graph's structure and properties.

The algorithm starts out with group formation, during which all nodes that have not been grouped yet are taken into consideration, in clustering-like fashion. In the initial run, two nodes with the maximum similarity of their neighborhood labels are grouped together. Their neighboring labels are modified to be equivalent immediately so that nodes in one group usually have the same neighbor labels.

For two nodes, v1 with neighborhood label set (LSv1), and v2 with neighborhood label set (LSv2), they calculated neighborhood label similarity (NLS) as follows:

$$NLS(v1, v2) = \frac{|LS_{V1} \cap LS_{V2}|}{|LS_{V1} \cup LS_{V2}|}$$

Greater value indicates greater similarity of the two neighborhoods.

Then nodes having the greatest similarity with any node in the group are clustered into the group till the group has l nodes with different sensitive labels.

Thereafter, the algorithm proceeds to create the next group. The residue nodes are clustered into existing groups as per the similarities between nodes and groups if fewer than l nodes are left after the last group's formation.

Identity disclosure of an individual and disclosure of sensitive labels is secured. An algorithm is applicable on web-based online social networking like face book, twitter. Facebook dataset is used.

ADVANTAGE: An algorithm not only hides the identity of users but also the selected features in users' profiles. The user can choose which features of her profile she wants to conceal.

DISADVANTAGE: Algorithm Direct Noise Node (DNN) is faster than the other two algorithms.

**2.2. Privacy Protection of Social Network Graphs [6]:**

Lijie Zhang and Weining Zhang [6] considered a vertex re-identification attack that partitions a social graph according to the neighborhoods of nodes. The neighborhood of a node will include direct neighbors of the vertex and the edges among these neighboring nodes. Their anonymization algorithm uses edge addition/deletion to construct a graph in which every vertex has the same neighborhood (in terms of structural isomorphism) as at least k-1 other nodes. This method uses the number of edges altered during anonymization as a measure of utility loss, and attempt to make as little edge change as possible.

Authors proposed Degree-Based Edge Algorithm. Degree-Based Edge Algorithm is a general framework for a degree-based edge anonymization. It takes as input a graph and a confidence threshold t, and returns a t-confident graph. The goal of the algorithm is to find a graph that not only satisfies the privacy requirement but also has a good utility. To achieve this goal, the algorithm uses a greedy strategy to improve graph confidence, that is it focuses on reducing the size of the leading EEC (one has the maximum linking probability) by perturbing the edge contained in the EEC. Intuitively, reducing the size of the leading EEC may improve graph confidence more quickly than reducing the size of other EECs, thus result in fewer edges being perturbed and better utility of anonymous graphs.

There are four strategies for anonymizing a graph by adding or deleting edges, namely, a) addition- only, b) deletion-only, c) swap-only and d) general addition/deletion. These strategies have different impact on the anonymization process. They also have different impact on various graph measurements (da F. Costa et al., 2007). For example, the swap-only does not change degree distribution, but may change the centrality and the shortest paths.

Identity disclosure of an individual and link disclosure is protected. It is applicable on Facebook, epinions. EPINION dataset is used.

ADVANTAGE: In this paper, the proposed method elaborated on a privacy measure for edge anonymity of an unlabeled, undirected social graph and edge anonymization algorithms based on this privacy measure using degree based graph partition. These algorithms alter social graphs by edge swap and edge deletion.

DISADVANTAGE: A clear understanding of various attacks is the basis of methods that protect privacy. The attacks discussed in this paper are mainly on social graphs.

**2.3. Personalized Privacy Protection in Social Networks [7]:**

M. Yuan, L. Chen, and P. S. Yu [7] focused on the privacy protection problem for an un-weighted graph with labels on both vertices and edges. Each vertex in the graph has different labels, which correspond to the attributes of the vertex. Every edge in the graph has one label that corresponds to the type of the edge. As one node may have distinct labels, they call the labels on node u as u's label list. They used G (V, E) to simply represent the original graph where V represents the node set and E represents the edge set.

Authors proposed a comprehensive privacy protection framework**.** This framework provides privacy protection services based on the user's personalized privacy protection requirements. Specifically, they defined three

levels of privacy protection requirements based on the moderately increasing adversary's background knowledge and merge the label generalization protection and the structure protection techniques (i.e. adding noise edge or nodes) simultaneously to fulfill different user's privacy protection demands.

This framework guaranteed the following protection objectives:

Given a constant k,

1. For every vertex u, the chances that an adversary re-identifies u is atmost $1/k$. For an adversary, re-identifying u is to discover which vertex is u in the release network using certain background knowledge regarding u;

2. For any edge e in the released network, the chances that an adversary identifies a vertex $u_x$ involved in e is atmost $1/k$.

3. For any two nodes $u_x$ and $u_y$, the chances that an adversary identifies these two vertices having a connection is atmost $1/k$.

For example, the probability that an adversary concludes Tim and Tom have a connection from the released graph should be less than or equal to $1/k$.

An algorithm protects the disclosure of node label and edge label representing type of relations. It is applicable onFacebook, Linkedin, and Livespace. Real datasets used: Speed Dating Data, ArXiv Data. Synthetic dataset used: ArXiv Data with uniform labels.

ADVANTAGE: The personalized privacy protection does not introduce extra computation time.

The algorithms introduced in this paper have good time efficiency.

DISADVANTAGE: The neighborhood graphs of the noise nodes do not have any special characters to be filtered out.

### 2.4. Identity Anonymization on Graphs [8]:

K. Liu and E. Terzi [8] considered a vertex re-identification attack that uses vertex degree to partition a social network graph. Their anonymization method uses edge deletion or edge addition to construct a k-anonymity graph that preserves the vertex degree distribution of the original graph.

Authors proposed Supergraph, Priority and GreedySwap algorithms. They proposed a two-step approach for the Graph Anonymization problem and its relaxed version. For an input graph G (V; E) with degree sequence d and an integer k, they proceed as follows:

1. First, starting from d, they construct a new degree sequence $d$ that is k-anonymous and such that the degree-anonymization cost $D_A(d,d) = L_1(d-d)$ is minimized.

2. Given the new degree sequence $d$, authors then construct a graph $G(V,E)$ such that $d_G = d$ and $E \cap E = E$ (or $E \cap E \approx E$ in the relaxed version).

Note that step 1 requires $L_1(d-d)$ to be minimized, which in fact translates into the requirement of the minimum number of edge additions due to Equation (1). Step 2 tries to construct a graph with degree sequence $d$, which is a supergraph (or has large overlap in its set of edges) with the original graph. If $d$ is the optimal solution to the problem in Step 1 and Step 2 outputs a graph with degree sequence $d$, then the output of this two-step process is the optimal solution to the Graph Anonymization problem.

Although in reality obtaining the optimal solution is not that easy, authors show how to solve the Graph Anonymization and its relaxed version by performing Steps 1 and 2 as described above. These two steps give rise to two problems, which they formally define. Performing step 1 translates into solving the Degree Anonymization problem. Authors proposed a priority algorithm for solving the degree anonymization problem. Similarly, performing step 2 translates into solving the Graph Construction problem. Authors proposed a probing scheme and greedy_swap algorithm for solving graph construction problem.

Identity disclosure of an individual is protected. Algorithm is applicable on any social networking dataset available.

Real datasets used: Co-authors graph, powergrid data. Synthetic datasets used: small-world graphs, scale-free graphs.Value of the exponent of the power-law distribution of the original and the k-degree anonymous graph is obtained using Supergraph, Priority and GreedySwap algorithms.

ADVANTAGE: An algorithm prevents the re-identification of individuals by an attacker with certain prior knowledge of the degrees.

DISADVANTAGE: An algorithm is not aware of any effective metrics to quantify the information loss incurred by the changes of its nodes and edges.

### 2.5. Privacy-Preserving Social Network Publication against Friendship Attacks [9]:

In a friendship attack, an attacker exploits the degrees of two nodes connected by an edge to re-identify related victims in a released social network data set. To secure against such thrashes, they introduced the concept of $k^2$-degree anonymity that limits the chances of a node being re-identified to 1/k. For $k^2$-degree anonymization problem, Tai, Yu, Yang and Chen [9] proposed an Integer Programming formulation to obtain optimal solutions in small-scale social networks. They proposed a scalable algorithm, called DEgree SEqence Anonymization (DESEAN), for $k^2$-degree anonymization of large-scale social networks.

For a published social network $\overline{G}$ of G, authors defined a friendship attack as follows:

For a target individual A and the degree pair information $D^2$ = (d1, d2), a friendship attack ($D^2$, A) uses $D^2$ to identify a vertex v1 that represents A in $\overline{G}$, in such a way that v1 joins to another node v2 in $\overline{G}$ with the degree pair $(d_{v1}, d_{v2})$ = (d1, d2).

The published social network $\overline{G}$ may have multiple candidate nodes that fulfill the above degree pair requirement. But, it is easy for an attacker to recognize A from the candidate nodes when the number of candidate nodes is small. Thus, to achieve privacy preservation, author defined $k^2$-degree anonymity as: A graph $\overline{G}$ is $k^2$-degree anonymous if, for each node with an incident edge of degree pair (d1, d2) in $\overline{G}$, there exist at least k – 1 other nodes, in such a way that each of the k – 1 vertices also has an incident edge of the same degree pair.

Algorithm DESEAN consists of three steps. The first step clusters vertices with similar degrees, chooses a target degree for every cluster, and ensures that each cluster includes at least k vertices. In order to achieve the required level of anonymity protection between two clusters, the second step adds or re- moves edges as required. The last step adjusts the edges in the graph such that all the vertices in each cluster meet the target degree selected in step 1.

Identity disclosure of an individual is protected. An algorithm is applicable on Facebook, MySpace or Friendster. Real datasets used: PODS09& 20TopConf. Synthetic datasets used: SD-SG & LS-SG from R-MAT graph model.

ADVANTAGE: An algorithm identifies the privacy risk in published social networks in terms of a new type of attack, called a friendship attack.

DISADVANTAGE: Only specific type of attack is detected.

### 2.6. Structural Re-identification in Anonymized Social Networks [10]:

Hay, Miklau, Jensen, Towsley & Weis [10] proposed the anonymization algorithm. They introduced a parameterized model of structural knowledge available to the adversary and quantify the success of attacks on people in anonymized networks. They show that the risks of these attacks vary based on network structure and size, and gives theoretical results that explain the anonymity risk in casual networks. They proposed a novel approach to anonymizing network information that models aggregate network structure and allow analysis to be performed by sampling from the model. The approach assures anonymity for entities in the network while allowing accurate estimates of a variety of network measures with relatively little bias.

Authors modeled a network as an undirected graph G = (V, E). The naive anonymization of G is an isomorphic graph, $G_a = (V_a, E_a)$ defined by a random bijection $\prod : V \to V_a$. For example small network represented as a graph along with its naive anonymization.

The anonymization mapping Π is a random, secret mapping. Naive anonymization prevents re-identification when an attacker has no information about individuals in the original graph. Formally stated, user x∈ V, called the target, has a candidate set, represented by cand(x), which contains the nodes of $G_a$ that could feasibly correlate with x. To

assess the risk of re-identification, they examine every element of the candidate set is equally likely and use the size of the candidate set as a measure of resistance to re-identification. As Π is random, in the lack of other information, any vertex in $G_a$ could related to the target vertex x. So, given an uninformed adversary, each user has the same risk of re-identification, specifically cand(x) = $V_a$ for every target node x.

But, if the adversary has access to external information about the entities, he could reduce the candidate set and threaten the privacy of individuals.

Set of articles connected by citations is protected; a communication network might describe Internet hosts related by traffic flows. Any social networking describes individuals connected by friendships. Real datasets used: HepTh, Enron, NetTrace. Synthetic datasets used: HOT, Power-Law, Tree & Mesh.

ADVANTAGE: Wide range of important graph analyses can be performed accurately on the generalized graphs published.

DIASADVANTAGE: Cannot develop bounds on the distortion introduced by anonymization.

**TABLE 1: C**omparison between Existing Methods for Privacy Protection on Social Networks

| *Sr. No.* | *Methodology* | *Parameters* | | | |
|---|---|---|---|---|---|
| | | *Proposed method* | *Data to be protected* | *Dataset used* | *Social network* |
| 2.1. | Sensitive Label Privacy Protection on Social Network Data[4] | Privacy protection algorithm | Identity disclosure and disclosure of node labels | Facebook dataset | Facebook, Twitter |
| 2.2. | Privacy Protection of Social Network Graphs[5] | Degree-Based edge algorithm | Identity disclosure and link disclosure | Epinion dataset | Facebook, epinions |
| 2.3. | Personalized Privacy Protection in Social Networks[6] | Comprehensive privacy protection framework | Disclosure of node label and edge label representing type of relationship | Real datasets:<br>Speed Dating Data, ArXiv Data<br>Synthetic dataset:<br>ArXiv Data with uniform labels | Facebook, Linkedin, and Livespace |
| 2.4. | Identity Anonymization on Graphs[7] | Supergraph, Priority and GreedySwap algorithms | Identity disclosure | Real datasets:<br>Co-authors graph , powergrid data<br>Synthetic datasets:<br>small-world graphs, scale-free graphs | - |
| 2.5. | Privacy-Preserving Social Network Publication Against Friendship Attack[8] | DEgree SEqence ANonymization (DESEAN) algorithm | Identity disclosure | Real datasets:<br>PODS09 &20TopConf<br>Synthetic dataset: SD-SG& LS-SG from R-MAT graph model | Facebook, MySpace or Friendster |
| 2.6. | Structural Re-identification in Anonymized Social Networks[9] | Anonymization algorithm | Link disclosure | Real datasets:<br>HepTh, Enron, NetTrace<br>Synthetic datasets:<br>HOT, Power-Law, Tree, Mesh | - |

## 3. Conclusion

In this paper we have reviewed different existing models and techniques for privacy protection of user sensitive data on social networks. By analyzing the existing system we will propose a privacy protection scheme that not only hides the identity of users but also the selected features in user's profiles.

## REFERENCES


[1] http://en.wikipedia.org/wiki/Privacy

[2] Boyd and Ellison, "Social Network Sites: Definition, History, Scholarship," 13 Journal of Computer-Mediated Communications 210, 2008.

[3] L. Liu, J. Wang, J. Liu, and J. Zhang, "Privacy preserving in social networks against sensitive edge disclosure," SIAM International Conference on Data Mining, 2009.

[4] Song, Panagiotis Kara's, Qian Xiao, and StéphaneBressan:"Sensitive Label Privacy Protection on Social Network Data, "IEEE Transactions on knowledge and data engineering, 25(3), 2013.

[5] S. Das, ¨O. Egecioglu, and A. E. Abbadi, "Anonymizing weighted social network graphs," ICDE, 2010.

[6] Lijie Zhang and Weining Zhang, "Privacy Protection of Social Network Graphs".

[7] M. Yuan, L. Chen, and P. S. Yu, "Personalized privacy protection in social networks," PVLDB, 4(2), 2010.

[8] K. Liu and E. Terzi, "Towards identity anonymization on graphs," SIGMOD, 2008.

[9] C.-H. Tai, P. S. Yu, D.-N. Yang and M.-S. Chen, "Privacy-preserving social network publication against friendship attacks," SIGKDD, 2011.

[10] Hay, M., Miklau, G., Jensen, D., Towsley, D., & Weis, P., "Resisting structural re-identification in anonymized social networks," International Conference on Very Large Data Bases.