Mining Social Media-Utility Based Privacy Preservation

V. Vijayalakshmi  
MCA Research Scholar  
Vels University  
Chennai, India

A.S. Arunchalam  
MCA , M.Phil  
Assistant Professor  
Vels University  
Chennai, India

R. Nandhakumar  
M.Tech.,(Phd)  
Xtream Technologies  
Chennai, India

Abstract-Online social networks and publication of social network data has led to the risk of leakage of confidential information of individuals. This requires the preservation of privacy before such network data is published by service providers. Privacy in online social networks data has been of utmost concern in recent years. Hence, the research in this field is still in its early years. Several published academic studies have proposed solutions for providing privacy of tabular micro-data. But those techniques cannot be straight forwardly applied to social network data as social network is a complex graphical structure of vertices and edges. Techniques like k-anonymity, its variants, L-diversity have been applied to social network data. Integrated technique of K-anonymity & L-diversity has also been developed to secure privacy of social network data in a better way.

General Terms- Social Network, Anonymization, Privacy, Attacks, Attributes.

Keywords- Privacy models, K-anonymity, L-diversity, t-closeness.

1. INTRODUCTION
Due to the increase in popularity of online social networks on the Web [1], large number of people subscribe to social networks or social media. This has generated large amount of user data that is gathered and maintained by the social network service providers. The data generated by social network services is termed as the social network data that needs to be published for others in certain situations. One of the situations is when specific analysis of the user data needs to be done and another situation is when the owner of the data has to share the data with third parties like advertising partners which is part of policies generally accepted by subscribers. The data contains valuable information about users that helps third parties in better social targeting of advertisements. Social network analysis is being used in modern sociology, geography, economics, and information sciences [2]. Researchers in various fields use this data for different purposes like researchers in government institutions require social network data for information and security purposes [3]. So, data needs to be shared or published in all above mentioned situations. Owner of data can publish it for others to analyze but it may create serious privacy threats. To fulfill the demands for the network data, online social media operators have been sharing the data they gather and maintain with external third parties such as advertisers, application developers, and academic researchers like Facebook has thousands of third-party applications and there has been an exponential increase in this number [4]. Social network data contains sensitive and confidential information about the users [5-7]. Thus sharing of this data in its raw form may breach privacy of individuals. Individual privacy is defined as “the right of the individual to decide what information about himself should be communicated to others and under what circumstances” [8]. A privacy breach occurs when private and confidential information about the user is disclosed to an adversary. So, preserving privacy of individuals while publishing user’s collected data is an important research area. Work has been done by various researchers in this direction.

This paper is structured as follows: Section 2 describes categories of privacy breach; followed by challenges in preserving privacy in social networks data which have been briefed in Section 3; Section 4 presents exiting techniques for preserving privacy in tabular micro-data; techniques for preserving privacy in social networks has been covered in Section 5; Section 6 gives research directions for new researchers; finally Section 7 concludes the review.

2. CATEGORIES OF PRIVACY BREACH
The privacy breaches in social networks can be categorized into three types [9-10]:

i. Identity disclosure - Identity disclosure occurs when an individual behind a record is exposed. This type of breach leads to the revelation of information of a user and relationship he/she shares with other individuals in the network.

ii. Sensitive link disclosure - Sensitive link disclosure occurs when the associations between two individuals are revealed. Social activities generate this type of information when social media services are utilized by users.

iii. Sensitive attribute disclosure – Sensitive attribute disclosure takes place when an attacker obtains the information of a sensitive and confidential user attribute. Sensitive attributes may be linked with an entity and link relationship.

All these mentioned privacy breaches pose severe threats like stalking, blackmailing and robbery because users expect privacy of their data from the service provider end. Besides that it damages the image and reputation of an individual. There are many examples of accidental disclosure of private information of users’ data that causes organizations to be conservative in releasing the network data, such as the AOL search data example [11] and attacks on Netflix data [12]. As per the promises of
social networks there is a need to address these issues. Therefore, data needs to be released to third publishing to third parties. But preserving privacy in social networks is difficult as mentioned in next section.

3. CHALLENGES IN PRESERVING PRIVACY IN SOCIAL NETWORK DATA PUBLISHING

Ensuring privacy for social network data is difficult than the tabular micro-data because [13]:

a) Modeling of background knowledge of adversaries is difficult in social network data than tabular micro-data. In tabular micro-data, users are identified by linking quasi-identifiers from whereas in social network information from various sources such as labels of vertices and edges, subgraphs, and neighborhood graphs can be used to identify individuals.

b) Information loss is the metric which measures the amount of distortion. In tabular micro-data information loss can be measured using the sum of information loss in individual records. Since, a social network is a graphical structure with a set of vertices and edges hence it is difficult to compare two social networks by comparing the vertices and edges individually. Anonymized social network and original social networks which have the same number of vertices and edges may have very different properties like betweenness, connectivity, and diameter. Information loss and anonymization quality can be measured in different ways.

c) Development of privacy preserving techniques in social network data is difficult than for relational data. Tabular micro-data is anonymized using divide-and-conquer techniques whereas social network is a structure of nodes.

5. PRIVACY PRESERVING TECHNIQUES – MICRO DATA

Significant work has been done for preserving privacy in tabular micro-data. Models like K-anonymity [14][20], L-diversity [15], T-closeness [16] have been proposed which have shown good results in anonymization. Fig. 1 briefs the three models, their properties and drawbacks.

Privacy preserving techniques are developed keeping following things into consideration:
1. Adversary’s knowledge
2. Utility of the data after release

So, depending upon the knowledge that an adversary uses to attack the target node following techniques have been developed by various researchers using the notion of K-anonymity [14][20]. Wei et al. [21] considered the privacy disclosure in online social network data publishing. It has been assumed that adversaries have the knowledge of the degree of a target individual and the target’s immediate neighbours. A practical solution to defend against background knowledge attacks has been proposed. Anonymized social networks obtained by proposed method can be used to answer aggregate network queries.
with high accuracy. Social network has been modeled as an undirected labeled graph. k-subgraph has been proposed to reduce the risk of privacy disclosure in social network data publication. Zou et al. [22] proposed k-automorphism based on the assumption that the adversary has knowledge about degree, subgraph and neighbor of the target node. Tripathy et al. [23] proposed an algorithm for graph isomorphism based on adjacency matrix. It says that a subgraph is indistinguishable from at least k-1 other subgraphs. Cheng et al. [24] used K-isomorphism to preserve privacy when adversary has subgraph knowledge. Wu et al. [25] proposed k-symmetry technique to protect privacy against re-identification using subgraph information. Lan et al. [26] developed an algorithm called KNAP against 1-pure structure anonymization methods. An anonymization methodology has been proposed by adding noise nodes into the original graph with the consideration of introducing the least distortion to graph properties. Skarkala et al. [27] applied K-anonymity to weighted social networks. Liu et al. [28] proposed the concept of k-degree to prevent vertex re-identification through the information of vertex degree.

Preserving privacy in social networks using k-anonymity protects against linking disclosure but still it may leak privacy under the cases of homogeneity and background knowledge attacks. Moreover, K-anonymity doesn’t protect against attribute disclosure. So, L-diversity was developed by Machanavajjhala [15] in year 2007.

Panda et al. [29] used a new practical and efficient definition of privacy called l-diversity on preserving privacy in collaborative social network data and the effect on the utility of the data for social network analysis has been seen. It has been identified that l-diversity social network still may leak privacy as an adversary may have some prior knowledge about the sensitive attribute value of an individual before seeing the released table. After seeing the released table, the adversary may have a posterior knowledge. Information gain i.e., the difference between the posterior knowledge and the prior knowledge is the factor to leak privacy. So the concept of t-closeness has been suggested to be introduced. Li et al. [30] proposed to preserve relationship privacy between two users one of whom can be identified in the released social network data. l-diversity anonymization model has been defined to preserve users’ relationship privacy. Two graph manipulation algorithms, MaxSub and MinSuper, have been proposed to achieve l-diversity anonymization.

Then, to preserve privacy in better way integrated approach of K-anonymity and L-diversity has been suggested by few authors as mentioned below.

Kavianpour et al. [31] proposed an integrated algorithm that takes the advantages of K-anonymity and L-diversity algorithm then evaluated the effectiveness of the combined strengths. Proposed algorithm has been able to increase the level of privacy for social network users by anonymizing and diversifying disclosed information. Tripathy et al. [32] proposed an algorithm which follows k-anonymity and l-diversity properties and can handle a variant of multisensitive attributes during anonymization process. Proposed algorithm is modified form of corresponding algorithms for micro data and it also depends upon some modified algorithms developed for anonymization against neighbourhood attacks. drawback of proposed algorithm is that it still needs some improvements in order to reduce the complexity so that it can be applied to large social networks. Yuan et al. [33] defined a k-degree l-diversity anonymity model for the protection of structural information and sensitive labels of people. Many privacy models like k-anonymity to prevent node reidentification through structure information have been proposed but an attacker may still be able to obtain private information of a person i.e. the label-node relationship is not well protected.

Other than above mentioned techniques for preserving privacy other techniques have also been proposed and developed as shown in table 1.

**Table 1. Various Other Privacy Preserving Techniques in Social Networks**

<table>
<thead>
<tr>
<th>Year</th>
<th>Author</th>
<th>Brief</th>
</tr>
</thead>
<tbody>
<tr>
<td>2008</td>
<td>Zhou et al. [34]</td>
<td>Reviewed existing anonymization techniques for privacy preserving publishing of social network data.</td>
</tr>
<tr>
<td>2008</td>
<td>Guha et al. [35]</td>
<td>Encryption has been used to provide privacy and only authorized users can decode and decrypt the result.</td>
</tr>
<tr>
<td>2008</td>
<td>Brosser et al. [36]</td>
<td>Proposed protocols to create and interact with privacy preserving collaborative social networks that combines small networks together while retaining the purity of data for the owners.</td>
</tr>
<tr>
<td>2008</td>
<td>Campan et al. [37]</td>
<td>Greedy approach to optimize utility using the attribute and structural information simultaneously has been used. Structural Information loss has been introduced. SANGREEA (Social Network Greedy Anonymization).</td>
</tr>
<tr>
<td>2008</td>
<td>Zheleva et al. [38]</td>
<td>How to preserve sensitive relationships.</td>
</tr>
<tr>
<td>2009</td>
<td>Ford et al. [39]</td>
<td>A new algorithm for enforcing p-sensitive k-anonymity on social network data based on a greedy clustering approach has been proposed.</td>
</tr>
<tr>
<td>2009</td>
<td>Narayanan et al. [40]</td>
<td>Developed Re-identification algo for anonymized graphs. Validated for Flickr and Twitter.</td>
</tr>
<tr>
<td>2009</td>
<td>Lijie et al. [41]</td>
<td>Studied link identification attack in which the adversary attacks using linking probability, t-confidence has been proposed. Dataset: EPINON, COA.</td>
</tr>
</tbody>
</table>
6. RESEARCH DIRECTIONS

Following are the few inferences drawn from literature survey:

1. To preserve usefulness/utility of anonymized data is an important aspect while applying techniques for privacy preservation. So, there is a need to develop methodologies that can quantitatively measure utility of data. There is need to evaluate various techniques in terms of tradeoff between privacy and utility.

2. Many algorithms like k-anonymity, L-diversity, integrated approach of k-anonymity & L-diversity have been developed for preserving privacy of social network user data but existing techniques leads to substantial information loss.

3. Anonymization techniques have been developed for one time released network data. But many applications require publishing data periodically so there is a need to develop techniques that can preserve privacy of dynamic releases.

4. Techniques are available for preserving privacy in case of distributed tabular data e.g. [59]. However, in

<table>
<thead>
<tr>
<th>Year</th>
<th>Reference</th>
<th>Summary</th>
</tr>
</thead>
<tbody>
<tr>
<td>2009</td>
<td>Ying et al. [42]</td>
<td>Considered edge re-identification attacks when the adversary has no background knowledge. Dataset: Enron, Email, Polblogs, Polbooks.</td>
</tr>
<tr>
<td>2009</td>
<td>Tootoonchian et al. [43]</td>
<td>Presented Lockr, a system that improves the privacy of centralized online system like Flickr and decentralized online content sharing systems like BitTorrent.</td>
</tr>
<tr>
<td>2009</td>
<td>Fong et al. [44]</td>
<td>Proposed an access control model that generalizes the privacy preservation mechanism of Facebook.</td>
</tr>
<tr>
<td>2010</td>
<td>Tang et al. [45]</td>
<td>Introduced KNN and EBB algorithm for constructing generalized subgraphs before sharing the social network with other parties and a mechanism to integrate the privacy preservation.</td>
</tr>
<tr>
<td>2009</td>
<td>Lan et al. [46]</td>
<td>Proposed an approach for preserving privacy of social networks which can be represented as bipartite graphs. Synthetic dataset.</td>
</tr>
<tr>
<td>2010</td>
<td>Ding et al. [47]</td>
<td>Presented a systematic review of the existing de-anonymization attacks in online social networks.</td>
</tr>
<tr>
<td>2010</td>
<td>Sun et al. [48]</td>
<td>Proposed a privacy-preserving method for sharing data in social networks with efficient revocation for preventing a contact's access right to the private data once the contact is removed from the social group and can be used as a plug-in for Facebook.</td>
</tr>
<tr>
<td>2010</td>
<td>Beach et al. [49]</td>
<td>q-Anon model has been presented to measure the probability of an attacker to identify unknown information from a social network API with the assumption that the data being protected may already be public. Validated on Facebook for 700 users.</td>
</tr>
<tr>
<td>2011</td>
<td>Yang et al. [52]</td>
<td>Non sensitive and generalized information has been used to support social network analysis and mining and to preserves the privacy of information.</td>
</tr>
<tr>
<td>2011</td>
<td>Zheleva et al. [53]</td>
<td>Surveyed the literature on privacy in social networks, Possible privacy breaches have been defined and possible privacy attacks have been studied.</td>
</tr>
<tr>
<td>2012</td>
<td>Fire et al. [54]</td>
<td>Developed Social Privacy Protector, software which aims to improve the security&amp;privacy of Facebook users.</td>
</tr>
<tr>
<td>2012</td>
<td>Masouzadeh et al. [55]</td>
<td>Proposed methods to enhance edge-perturbing anonymization on the basis of structural roles and edge betweenness in social network theory. Dataset: Polbook, Jazz</td>
</tr>
<tr>
<td>2013</td>
<td>Tassa et al. [56]</td>
<td>The first study of privacy preservation in distributed social networks which is shown to outperform SaNGreeA algorithm which is the leading algorithm for achieving anonymity in networks by means of clustering.</td>
</tr>
<tr>
<td>2013</td>
<td>Heathely et al. [57]</td>
<td>Examined that friendship links and details altogether provide better predictability than details alone, effect of removing details and links in preventing sensitive information leakage has been explored.</td>
</tr>
<tr>
<td>2013</td>
<td>Cheng et al. [58]</td>
<td>Proposed a framework to provide users controls over how third party applications can access their data and activities in social networks while still retaining the functionality of third party applications.</td>
</tr>
</tbody>
</table>
case of social network distributed privacy preserving techniques are not well reported in literature except [56].  
5. Existing privacy preserving approaches for social networks have been evaluated using either small datasets or synthetic datasets. There is need to conduct empirical experiments on large datasets.  
6. There is no existing technique which can prevent homogeneity attacks, background knowledge attacks, attacks arising due to distance between sensitive values.

7. CONCLUSION

It became evident from the literature that privacy of users is the main concern and topic of research now a days. Various models proposed for tabular micro-data have been adopted for preserving privacy of social network data. Techniques like K-anonymity, L-diversity, integrated K-anonymity L-diversity have been used till now but these techniques lead to substantial information loss. So, there is a scope of improvement of the techniques that provide privacy preservation with minimum information loss and better utility of released data.

REFERENCES

of Social Networks (CASoN), Sao Carlos, pp 126 – 131, 2012.


