ISSN:0975-9646

P.Mohana Lakshmi et al, / (IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 5 (4) , 2014, 5647-5650

Sequential Clustering Algorithms for Anonymizing Social Networks

P.Mohana Lakshmi
M.Tech : Department of CSE
SIETK,Puttur,INDIA

P.Balaji M.Tech
Associate Professor
Department of CSE
SIETK,Puttur,INDIA

P.Nirupama M.E
Head of the Department
Department of CSE
SIETK,Puttur,INDIA

Abstract— The privacy-preservation in social networks is major problem in now-a-days. In distributed setting the complex data is divide between several data holders. The target is to appear at an anonymized view of the unified network without illuminating to any of the data holders information about links between nodes that are hold by other data holders. To that finish, in centralized setting two variants of an anonymization algorithm are offered which is based on sequential clustering (Sq). Proposed algorithms substantially break the SaNGreeA algorithm due to Campan and Truta which is the primary algorithm for achieving anonymity in networks by means of clustering and then secure distributed versions of algorithms. To the top of awareness, this is the earliest study of privacy preservation in distributed social networks. Finally, conclude by outlining potential research proposals in that path.

Keywords—social networks, clustering, privacy preservation, distributed computation

I. RELATED WORK

Now-a-days the use of social networks among the people has become more popular. With the impact of social networks on society, the people become more sensitive regarding privacy issues in the common networks and most sociologists agree that this trend will not fade away. Privacy preservation in social networks is major problem in now-a-days. This problem can be solved as explained follows.

The advent of social network sites in the last years seems to be a trend that will likely continue. What naive technology users may not realize is that the information they provide online is stored implications of massive data gathering, and effort has been made to protect the data from unauthorized disclosure. However, the data privacy research has mostly targeted traditional data models such as micro data. Recently, social network data has begun to be analyzed from a specific privacy perspective, one that considers, besides the attribute values that characterize the individual entities in the networks, their relationships with other entities. Campan and Truta[1] proposed a greedy algorithm for anonymizing social network and a measure that quantifies the information loss in the anonymization process due to edge generalization.

Publishing data about individuals without revealing sensitive information about them is an important problem. In recent years, a new definition of privacy called k-anonymity has gained popularity. In a k-anonymized dataset, each record is indistinguishable from at least k − 1 other records with respect to certain "identifying" attributes. Two simple attacks that a k-anonymized dataset has some subtle, but severe privacy problems. First, an attacker can discover the values of sensitive attributes when there is little diversity in those sensitive attributes. This is a known problem. Second, attackers often have background knowledge, and A.Machanavajjhala, D.Kifer, J.Gehrke and M.Venkitasubramaniam [2] show that k-anonymity does not guarantee privacy against attackers using background knowledge. They give a detailed analysis of these two attacks and propose a novel and powerful privacy criterion called ℓ-diversity that can defend against such attacks. In addition to building a formal foundation for ℓ-diversity, they show in an experimental evaluation that ℓ-diversity is practical and can be implemented efficiently.

One of the most well studied models of privacy preservation is k-anonymity. Previous studies of k-anonymization used various utility measures that aim at enhancing the correlation between the original public data and generalised public data. J.Goldberger and T.Tassa [3] bearing in mind that a primary goal in releasing the anonymized database for data mining is to deduce methods of predicting the private data from the public data, and propose a new information-theoretic measure that aims at enhancing the correlation between the generalised public data and private data. Such a measure significantly enhances utility of the released anonymized database for data mining. They proceed to describe a new algorithm that is designed to achieve k-anonymity with high utility, independently of the underlying utility measure .that algorithm is based on a modified version of sequential clustering which is the method of choice in clustering. Expreimental comparison with four well known algorithms of k-anonymity show that the sequential clustering algorithm is an efficient algorithm that achieves the best utility results. They describe a modification of the algorithm that outputs k-anonymizations which respect the additional security measure of l-diversity.

Consider the distributed setting in which the network data is split between several data holders. The goal is to arrive at an anonymized view of the unified network without revealing to any of the data holders information about links between nodes that are controlled by other data holders .Algorithms significantly outperform SaNGreeA algorithm due to Campan and Truta which is the leading algorithm

for achieving anonymity in networks by means of clustering. Tamir Tassa and Dror J.Cohen [4] planned secure distributed versions of algorithms. Those algorithms produce anonymizations by means of clustering better utility than those achieved by existing systems. The goal of the proposed work is to arrive at an anonymized view of the social network without revealing to any of the data holders information about the nodes and links between nodes that are controlled by data holders.

II. INTRODUCTION

Networks are structures that describe a set of entities and the relations between them. A social network, for example, provides information on individuals in some population and links between them [5].In their most basic form, networks are modeled by a graph where the nodes and edges corresponds to entities and their relationships between them. Real social network may be more complex or may contain additional information. Hence, it is modeled as a hyper-graph. When there are several types of interactions indulged, then the edges would be labeled, or the graph could be accompanied by attributes. Data in social network need to be anonymized before its publication in order to preserve the privacy of individuals by concealing sensitive information.

A naive anonymization of the network by removing the identifiable attributes like names, zip code, etc., from the data is inadequate. The theme behind the attack [6] is to inject a group of nodes with a distinctive pattern of edges among them in the network. The adversary links the patterns and the targeted node is subjected to attack.

III. EXISTING SYSTEM

The existing system suffers issues related to privacy. The data in such social network cannot be released as it is, since it might contain sensitive information. As predicted earlier, a naive anonymization of removing identifying attributes is insufficient. Hence a more substantial procedure of anonymization is required. The methods of privacy preservation in the existing system can be well defined by means of three categories.

- The first category provides k-anonymity via deterministic procedure of edge additions or deletions.
- The second category adds noise to the data, in the form of a random additions, deletions or switching of edges.
- The third category don't follow the method of altering graphs, instead they cluster together nodes into super nodes.

Limitations of existing system:

- The study of anonymizing social networks has more concentrated so far on centralized networks only.
- Privacy cannot be maintained thoroughly since every single detail is visible to all.
- A naive anonymization is insufficient. It is possible to collect information from a social graph in an efficient manner.
- The premise of collecting and analyzing information from a user's explicit or implicit social network enhances the accuracy rate of search results

IV. PROPOSED SYSTEM

Though, the exiting categories of privacy preservation is good, so far concentrated only on centralized networks and more over the existing technique still holds some issues of security and privacy breeches. To tackle such constraints, the proposed algorithm issues anonymized views of graph with significantly smaller information losses than anonymization techniques issued by earlier algorithm. These works stays in the realm of network and propose two variants of an anonymization algorithm which is based on sequential clustering. A distributed version of this algorithm computes a kanonymization of the unified network by invoking secure multiparty protocols.

A. The Data

The social network is viewed as a simple undirected graph is G= (V, E), where V= {v1,.....,vN} is the set of nodes and E$\underline{c}$(v_2)is the set of edges. Each node corresponds to an individual in the underlying group, while an edge describes the relationships among nodes by connecting them. Non-identifying attributes are called quasi-identifiers. For example age, zip code, etc.,. To that linking attacks [7] quasi-identifiers are used

B. Anonymization by clustering

Anonymization of given social network is done SN=(V, E, R) by means of clustering as predicted in [1],[8],[9]. Given a clustering C= {c1 ...cT} of v, which are the clusters or disjoint subsets. The corresponding clustered Social network is SNC=(C, EC,).The clusters are labeled by their size and number of inter-cluster edges .Given social network SN= (V, E, R) a corresponding clustered social network is called K-anonymous or K-anonymization of social network if the size of all its clusters is atleast k.

C .Measuring the loss of information

The measuring techniques are inherited from [1]for the analysis of information loss in the considered social network. Given a social network and a clustering C of its nodes, the information loss associated with replacing social network by corresponding SNC is defined as a weighted sum of two metrics.

$$I(c) = w.ID(c) + (1-w).IS(c)$$

Here, w_[0,1] is some weighing parameter, ID(C) is the descriptive information loss & IS(C) is the structural information loss. For the descriptive metric, the Loss Metric (LM) measure is utilized from [10] [11]. The structural information loss is classified as Intra-Cluster information loss & Inter-Cluster information loss. All the loss measures range between 0 & 1.

D. Previous Algorithm of K-Anonymization by Clustering

The first anonymization algorithm by taking account of both descriptive & Structural data was SANGreeA [1].But it suffers the problem of Structural information loss when clustering of nodes attains K-Anonymity. But the presented Sequential clustering algorithm doesn't suffer such problem. In each stage of its execution it has a full clustering which prevents the information loss measure.

V. PROPOSED TECHNIQUES

A. Anonymization by Sequential Clustering

K-Anonymization of tables using sequential clustering Mechanism is dealt in [10]. It was shown that, it's the efficient technique in terms of runtime as well as is terms of utility of the output anonymization. This technique avoids the loss of information, for example: if we have a huge number of data means the grid view size of thedata is enlarged. This proceeds with an adoption which starts with a random portioning of the network nodes into clusters. Then, the nodes are moved in a cyclic manner for checking whether that node may be moved from its current cluster to another one while decreasing the information loss of the induced anonymization. If such an improvement is possible, the node is transferred to the cluster where it currently fitbest.

A Modified Structural Information loss measure

The proposed SANGeerA algorithm [9] uses a measure of structural information loss that differs from the measure of actual information loss. Since, it is defined as a sum of independent intra-cluster information loss measures. As the SANGreeA algorithm needs to make clustering decision before all clusters are formed, it uses a distance for between a node & a cluster that's geared towards minimizing the measure of structural information loss.

B. Distributed Setting

There are 2 scenarios to consider in this setting:

- Scenario A: Each player (peers) needs to protect the identifier of the nodes under his control from other players, as well as the existence or non-existence of edges adjacent to his nodes.
- Scenario B: All players (peers) know the identifier of all nodes in the vertex; the information that each player needs to protect from other players is the existence or nonexistence of edges adjacent to his nodes.

The analysis of distributed setting is described by the analysis of Distributed Sequential Clustering & implementation of distributed & centralized network with primary by decreasing the limitations of Kanonymity algorithm & communication complexity.

VI. ALGORITHM DESCRIPTION

The sequential clustering algorithm for k-anonymizing tables was presented in [1]. It was shown there to be a very efficient algorithm in terms of runtime as well as in terms of the utility of the output anonymization. We proceed to describe an adaptation of it for anonymizing social networks.

The algorithm starts with a random partitioning of the network nodes into clusters. The initial number of clusters in the random partition is set to $\lfloor N/k0 \rfloor$ and the initial clusters are chosen so that all of them are of size $k0$ or $k0 + 1$, where $k0 = \alpha k$ is an integer and α is some parameter that needs to be determined.The algorithm then starts its main loop (Steps 2-4). In that loop, the algorithm goes over the N nodes in a cyclic manner and for each node it checks whether that node may be moved from its current cluster to another one while decreasing the information loss of the induced anonymization. If such an improvement is possible, the node is transferred to the cluster.

Algorithm:

Input: A social network SN, an integer k.

Output: A clustering of SN into clusters of size $\geq k$.

1)Choose a random partition C = {C1, . . . ,CT } of V Into T := [N/k0] clusters of sizes either k0 or k0 + 1.

2) For $n = 1, \ldots, N$ do:

a) Let Ct be the cluster to which vn currently belongs.

b) For each of the other clusters, Cs, $s \neq t$, compute the difference in the information loss, $\Delta n{:}t \rightarrow s$, if vn would move from Ct to Cs.

c) Let $Cs0$ be the cluster for which $\Delta n{:}t \rightarrow s$ is minimal.

d) If Ct is a singleton, move vn from Ct to $Cs0$ remove cluster Ct.

e) Else, if $\Delta n{:}t \rightarrow s0 < 0$, move vn from Ct to $Cs0$.

3) If there exist clusters of size greater than $k1$, split each of them randomly into two equally-sized clusters.

4) If at least one node was moved during the last loop, go to Step 2.

5) While there exist clusters of size smaller than k, select one of them and unify it with the cluster which is closest.

6) Output the resulting clustering.

During that main loop, we allow the size of clusters is to vary in the range $[2, k1]$, where $k1 = \beta k$ for some predetermined fixed parameter β. When a cluster becomes a singleton, remove it and transfer the node that was in that cluster to the cluster where it fits best, in terms of information loss (Step 2d). On the other hand, when a cluster becomes too large (i.e., its size becomes larger than the upper bound $k1$),we split it into two equally-sized clusters in a random manner. the main loop of the algorithm is repeated until we reach a stage where an entire loop over all nodes in the network found no node that could be moved to another cluster.

CONCLUSION

Sequential clustering algorithms for anonymizing social networks are presented. Those algorithms can produce anonymization by means of clustering with better utility than those achieved by existing algorithms. A secure distributed version of this algorithm for the case in which the network data is split between several nodes is devised. We focused on the scenario in which the interacting peers know the identity of all nodes in the network, but need to protect the structural information(edges) of the network. In this scenario, each of the peers needs to protect the identity of the nodes under his control from the other peers. Hence, it is more difficult in two manners: It requires a secure computation of the descriptive information loss (while in existing such a computation can be made in a public manner); and the peers must hide from other peers the allocation of their nodes to clusters.

REFERENCES

[1] A. Campan and T. M. Truta. Data and structural k-anonymity in social networks. In *PinKDD*, pages 33–54, 2008.

[2] A. Machanavajjhala, D. Kifer, J. Gehrke, and M. Venkitasubramaniam. *ℓ*-diversity: Privacy beyond *k*-anonymity. *ACM Trans. Knowl. Discov. Data*, 1(1):3, 2007.

[3] J. Goldberger and T. Tassa. Efficient anonymizations with enhanced utility. *TDP*, 3:149–175, 2010.

[4] Tamir Tassa and Dror J.Cohen, "Anonymization of centralized and distributed social networks by sequential clustering" IEEE Transactions on Knowledge and data Engineering, vol.25, no.2, 2013.

[5] M. Hay, G. Miklau, D. Jensen, P. Weis, and S.Srivastava. Anonymizing social networks. *Uni. Of Massachusetts Technical Report*, 07(19), 2007.

[6] L. Backstrom, C. Dwork, and J. M. Kleinberg. Wherefore art thour3579x?: anonymized social networks, hidden patterns, and structural steganography. In *WWW*, pages 181–190, 2007.

[7] L. Sweeney. Uniqueness of simple demographics in the U.S. population. In Laboratory for International Data Privacy (LIDAP-WP4), 2000.

[8] M. Hay, G. Miklau, D. Jensen, D. F. Towsley, and P.Weis. Resisting structural re identification in anonymized social networks. In *PVLDB*, pages 102–114, 2008.

[9] E. Zheleva and L. Getoor. Preserving the privacy of sensitive relationship in graph data. In *PinKDD*, pages 153–171, 2007.

[10] V. Iyengar. Transforming data to satisfy privacy constraints. In *ACMSIGKDD*, pages 279–288, 2002.

[11] M. E. Nergiz and C. Clifton. Thoughts on *k*anonymization. In *ICDE Workshops*, page 96, 2006.