Incognizance of Social Networks by Sequential Clustering

Ms.B.Gothai[1], Mr.V.Janaki Raman[2]

[1]PG Scholar, Department of Computer Science and Engineering, DR.Pauls Engineering College, Villupuram, Tamil Nadu-605 109

[2]Assistant Professor, Department of Computer Science and Engineering, DR. Pauls Engineering College, Villupuram, Tamil Nadu-605 109

Abstract

The complexity in preserving privacy of social networks is considered. The distributed setting in which the network data is split between several data holders is focussed. The goal is to arrive at an anonymized view of the unified network in a distributed environment. The leading clustering algorithm for achieving anonymity is SANGreeA (Social Network Greedy Anonymization), which is significantly outperformed by our proposed clustering algorithmic techniques. To the best of our knowledge, this is the first study for privacy preservation in distributed social network.

Index terms: *Social networks; clustering; privacy preserving data mining; distributed computation.*

I. 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 [1].In their most basic form, networks are modelled 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 modelled as a hyper-graph. When there are several types of interactions indulged, then the edges would be labelled, 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 [2 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.

II. 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.

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

Limitations of existing system

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

III. RELATED WORKS

Complex real time public networks are designed and manipulated through graph modulation. Names are replaced by meaningless unique identifiers Backstrom

[6]. Labels are unique and unidentifiable technique. Cryptography is the technique that is used for anonymization of networks. Enciphering and Deciphering plays a vital role in implementing this technique. Adversary links distinctive structure. Sub graph is tracked from naively anonymized network [2] Liu. K and Terzi .E. This definition of anonymity prevents the re-identification of individuals by adversaries with *a priori* knowledge of the degree of certain nodes. Formally, the graph-anonymization problem for a given graph *G*, asks for the *k*-degree anonymous graph that stems from *G* with the minimum number of graph-modification operations. A simple and efficient algorithm for solving this problem is devised. This work is based on principles related to the reliability of degree sequences. This method is applied to a large spectrum of synthetic and real datasets and demonstrates their efficiency and practical utility. The disadvantages are Capability of attacker is unknown, difficult to measure utility of graph.

IV. PROPOSED SYSTEM

Though, the exiting categories of privacy preservation is good, so far concentrated only on centralized networks and moreover the existing technique still holds some issues of security and privacy breeches. To tackle such constraints, the proposed algorithm issues anonymized views of the graph with significantly smaller information loses 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 k-anonymization of the unified network by invoking secure multiparty protocols.

4.1 The Data

The social network is viewed as a simple undirected graph, G= (V, E), where V= {$v_{1,......,}v_N$} is the set of nodes and E$\underline{\text{c}}\binom{V}{2}$ 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 thwart linking attacks [3] quasi-identifiers are used in combinations.

4.2 Anonymization by clustering

Anonymization of a given social network is done SN= (V, E, R) by means of clustering as predicted in [4], [5], [6]. Given a clustering C= {c_1 ...c_T} of v, which are the clusters or disjoint subsets. The corresponding clustered Social network is SN_C=(C, $E_{C,}$).The clusters are labelled by their size and number of inter-cluster edges. Given a 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.

4.3 Measuring the loss of information

The measuring techniques are inherited from [7] 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 SN_C is defined as a weighted sum of two metrics.

$$I(c) =w.I_D(c) + (1-w).I_S(c)$$

Here, wℇ[0,1] is some weighing parameter, I_D(C) is the descriptive information loss & I_S(C) is the structural information loss. For the descriptive metric, the Loss Metric (LM) measure is utilized from [8] [9]. The structural information loss is classified as Intra-Cluster information loss & Inter-Cluster information loss. All the loss measures range between 0 & 1.

4.4 Previous Algorithm of K-Anonymization by Clustering

The first anonymization algorithm by taking account of both descriptive & Structural data was SANGreeA [7]. 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.

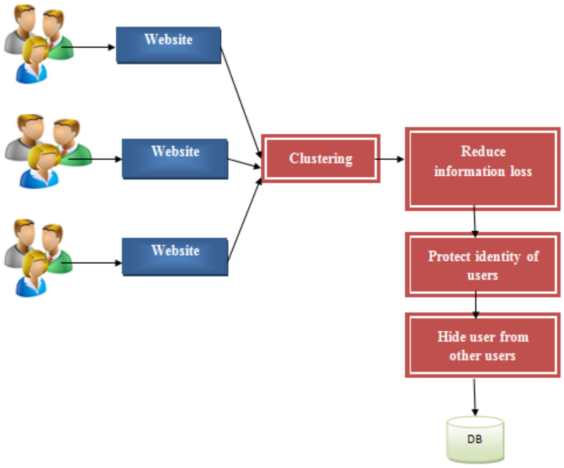

Figure1: System Model

V. PROPOSED TECHNIQUES

5.1 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 the data 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 fit best.

A Modified Structural Information loss measure

The proposed SANGeerA algorithm [6] 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.

5.2 Distributed Setting

There are 2 scenarios to consider in this setting:

1. 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.
2. 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 non-existence 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 K-anonymity algorithm & communication complexity.

VI. CONCLUSION

Sequential clustering algorithms for anonymizing social networks are presented. Those algorithms 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] M. Hay, G. Miklau, D. Jensen, P. Weis, and S. Srivastava. Anonymizing social networks. *Uni. of Massachusetts Technical Report*, 07(19), 2007.

[2] 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.

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

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

[5] 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.

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

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

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

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

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

[11] D. Watts and S. Strogatz. Collective dynamics of 'small-world' networks. *Nature*, 393:409–410, 1998.

[12] A. Barab´asi and R. Albert. Emergence of scaling in random networks. *Science*, 286:509–512, 1999.