Scientific Journal Impact Factor (SJIF) : 1.711
International Journal of Modern Trends in Engineering and Research
ISSN (ONLINE) : 2349 - 9745 ISSN (PRINT) : 2393 - 8161

ANONYMIZATION OF SOCIAL NETWORKS FOR REDUCING COMMUNICATION COMPLEXITY AND INFORMATION LOSS BY SEQUENTIAL CLUSTERING

Nirav. U.Patel[1], Vaishali.R.Patel[2]
[1]Department of Computer Engineering / IT,SVM Institute of Technology, Bharuch 392-001, Gujarat, India
[2]Department of Computer Engineering / IT,SVM Institute of Technology Bharuch 392-001, Gujarat, India

Abstract— Privacy Preserving Publishing in Social Network is always an important Concern. In Recent Years, the impact of social networks on society, the people become more sensitive regarding privacy issues in the Social networks. The goal of the proposed work is to arrive at an anonymized view of the social networks without revealing to any information about the nodes and links between nodes that are controlled by the data holders. The main contributions in this paper are Hierarchical algorithm for anonymizing a social network and a measure that quantifies the information loss in the anonymization process to preserve privacy. The anonymized dataset permits strong attacks due to lack of diversity in the sensitive attributes. This paper uses the t-closeness, a framework that gives stronger privacy guarantees.

Keywords— Social networks; Anonymization; Hierarchical clustering; Information Loss; t-closeness.

I. INTRODUCTION

With the rapid growth of social networks in society, more and more people participate in social networks. Social networks Model Social relationships by graph structure using vertices and edges. In Social Networks Vertices describe individual actor and edges describe relationships between Vertices (actor).Many different kind of Social networks present in our lives such as MySpace, Facebook, Twitter. India now has 243.2 million internet users and 106 million active social media users, according to the latest mid-year figures for 2014.

Social networks connect social actors. The connections are often beneficial to enterprises and commercial companies. For example, they can use the connections to expand their customer bases. In many cases, those social networks can serve as a customer relationship management tool for companies selling products and services. Companies can also use social networks to identify potential customers or recruit candidate employees. For example, according to the statistics published in Time India Magazine, 11% of employers in the India use popular social networking sites such as MySpace and Facebook to investigate potential employees. With the rapid growth of social networks, social network analysis [3] has emerged as a key technique in modern sociology, geography, economics, and information science. The goal of social network analysis is to uncover hidden social patterns. The power of social network analysis has been shown much stronger than that of traditional methods which focus on analyzing the attributes of individual social actors. Anonymization of the data can mitigate privacy and security concerns. Data anonymization may not take away the original field layout of data being anonymized and the data may be look realistic. Anonymization can be done by aggregation, hashing clustering, generalization, etc. It is used to increases the user privacy.

This work proposes a new anonymization approach which involves clustering, generalization, suppression approaches. It prevents the quasi-identifier information of the individual's from

disclosure. To prevent the individual's sensitive information from disclosure the privacy measure t-closeness is used in this work.
The rest of this paper is structured as follows. In section II we review the basic concepts of generalization and suppression, and then we discuss the calculations of information loss measures in section III. In section IV we survey the existing techniques. We describe our proposed Hierarchical clustering algorithm in section V. We give the overview of t-closeness in section VI. Experimental results are discussed in section VII. Finally, section VIII concludes this paper.

II. ANONYMIZATION BY CLUSTERING

In privacy preserving data publishing, in order to prevent privacy attacks, data should be anonymized properly before it is released. The primary goal in releasing the anonymized database for data mining is to deduce methods of predicting the private data from the public data [11]. This paper described the sequential clustering algorithm for k-anonymization. This algorithm starts by treating each data object as a cluster and then recursively merges these clusters based on minimum distance between them until all objects are in a single cluster. As there is no guarantee that such procedure finds the Networks are structures that describe a set of entities and the relations between them. 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 [2].This paper contributed the Sequential Clustering Algorithm which performs clustering process to generate k-anonymized Social network. The algorithm establishes random partitions of all nodes in to clusters. Then this paper also introduces a measure to quantify generalization and loss of information.
The publishing of Social network data about individuals without revealing sensitive information about them is major problem [5].This paper proposed novel approach called l-Diversity. K-Anonymity protects against identity disclosure, it does not provide sufficient protection against attribute disclosure. The notion of l-diversity attempts to solve this problem by requiring that each equivalence class has at least one well represented value for each sensitive attribute. 1-diversity is practical, easy to understand and addresses the shortcomings of k-anonymity with respect to the background knowledge and homogeneity attacks. But in some situation when multiple records in the table correspond to one individual it cannot prevent the attribute disclosure.

III. AGGLOMERATIVE HIERARCHICAL CLUSTERING

The process of grouping a set of data points into classes of similar objects is called clustering [4]. The fundamental idea of clustering is that the intra-cluster similarity should be high and the inter-cluster similarity should be low. This bottom-up strategy of clustering starts by treating each data object as a cluster and then recursively merges these clusters based on the minimum distance between them until all objects are in a single cluster or until a condition is satisfied. At each step, a cluster is represented by the mean value of all the objects in the cluster, i.e., the centroid of that cluster.

A tree structure called dendrogram is commonly used to represent the process of hierarchical clustering [4]. It is a step by step diagram of the clustering process where the clustering is represented as the fusion of branches of the tree.

Algorithm-I: Data Anonymization by Agglomerative Hierarchical Clustering

---

Input: Dataset D containing set of n objects.
Output: Cluster of objects.

Method:

1. Create 'm' clusters corresponding to each data object.
2. Repeat
3. Distance (C_p,C_q)=minimum(distance(C_i,C_j)), for all i , j forming cluster pairs.
4. Merge the clusters C_p and C_q.
5. Update the cluster centroid of merged cluster.
6. Repeat steps 3 and step 4 .
7. Until, only 1 cluster remains.

IV. GENERALIZATION AND SUPPRESSION

Generalization means replacing quasi-identifiers [1], the original value is replaced by some more specific to less value, such as age. For example the Age 25 can be generalized to $20 < \text{Age} \leq 30$. Various generalization strategies have been proposed in [6, 13, 14]. Those papers allow value from different domain levels to be combined to generate the generalization. Generalization on graph data can be done by using one of the five categories proposed in [7].

Suppression replace the original value by some specific symbols. For example the gender {Male, Female} attribute is completely suppressed to *. It can also be done by removing individual attribute values or by partitioning the attribute domain into intervals.

V. INFORMATION LOSS MEASURES

There are two information loss measures, Generalization information loss measure and Structural information loss measure.

The generalization information loss measure describes how much descriptive data are lost during generalization. The structural information loss measure describes how much structural details are lost during generalization. The generalization information loss measure was introduced in [2, 11, 12, 15]. In this paper, the generalization information loss measure is used as described in [2]. Let Ca be a cluster, n is the number of generalizations and p(i) be the size of each generalized data and a[i] be the number of the data in each attribute before generalization. Then the generalization information loss measure

GI is defined as

$$GI(C_a) = \frac{1}{n} \cdot \sum_{i=1}^{n} \frac{|p(i)| - 1}{|a(i)| - 1}$$

The total Generalization Information Loss is measured by

$$TG = \frac{1}{N} \cdot \sum_{i=1}^{c} Ki \cdot GI$$

Where;

N - Number of records in the table.

K_i - The size of Cluster.

GI- The generalization information loss.
C - The number of clusters.
When anonymizing a graph some of the structural informations will be lost, that can be measured using structural information loss measure. There are two types of structural information losses. Intra cluster information loss and Inter cluster information loss. Given a cluster Ca, $1 \leq a \leq T$, the original graph is replaced by the number of nodes Ka in the cluster Ca and number of edges Ea in the cluster Ca. Then the intra cluster information loss is measured as

$$IA(C_a) = 2Ea.\left(1 - \frac{2Ea}{|Ka|.(|Ka|-1)}\right)$$

Given two clusters Ca and Cb the structure of the edges that
connects the nodes in Ca to the nodes in Cb are lost by replacing it by the number of edges between the nodes in those two clusters. Then the inter cluster information loss is measured as

$$IR(C_a) = 2E_{ab}.\left(1 - \frac{2Eab}{|Ka|.(|Kb|)}\right)$$

The total Structural information loss is calculated by using the following formula

$$SI(C) = \frac{4}{N(N-1)}\left(\sum_{t=1}^{T} IA(C_a) + \sum_{1 \leq a \neq b \leq T} IR(C_a)\right)$$

Where SI(C) ranges between zero and one.

VI. T-CLOSENESS: A PRIVACY MEASURE

Privacy is measured by the information gain of an observer. The gain is the difference between the prior belief and the posterior belief. A novel privacy measure t-closeness is proposed in [9]. It requires the distance between the distributions of a sensitive attribute and the distribution of the attribute in whole table should be no more than the threshold t. The distance between two probabilistic distributions can be measured using Earth Mover's Distance (EMD) [8]. EMD requires that the distance between the two probabilistic distributions to be dependent upon the ground distances among the values of an attribute.
For Numerical Attributes the EMD is calculated as follows. Let $X=(x_1,x_2,\ldots,x_n)$, $Y=(y_1,y_2,\ldots,y_n)$ be the given two distributions and dij be the ground distance between the element i of X and element j of Y. If the element 1 has an extra amount of x_1-y_1, then the amount of y_1-x_1 should be transported from other elements to the element 1. And also, element 1 is transported from element 2. After that element 1 is satisfied and element 2 has an extra amount of $(x_1$-$y_1)$ +$(x_2$-$y_2)$. The above described process continues until element m is satisfied and Y is reached. Let $r_i = x_i$-y_i, (i=1,2,..,n) then the distance between X and Y can be calculated as

$$D[X,Y] = \frac{1}{n-1}\left(|r_1|+|r_1+r_2|+\ldots+|r_1+r_2+r_{n-1}|\right)$$

$$= \frac{1}{n-1}\sum_{i=1}^{n}\left|\sum_{j=1}^{i} r_j\right|$$

For Categorical attributes the ground distance between any two values of a categorical attribute is 1. If the distance between any two values is 1 then one value should be moved to some other points.

$$D[X,Y] = \frac{1}{2}\sum_{i=1}^{n} |x_i - y_i|$$

T-closeness overcomes the background knowledge attack and similarity attack which is the drawback of k-anonymity [5]. It protects against attribute disclosure. In this work we use both t-closeness and k-anonymity using Hierarchical clustering at the same time for better results.

VII. EXPERIMENTAL RESULTS

This section compares the performance of the proposed Agglomerative Hierarchical algorithm including t-closeness privacy measure against the combination of sequential clustering algorithm and l-diversity privacy measure. Both algorithms are implemented in Matlab 2012b and run on a w Intel coreTM i5 processor, 4GB of RAM under the Windows 8 operating system. This experiment uses the Census dataset; that data consist of Seven attributes including age, gender, education, work class, zipcodes, race and country are quasi identifiers and then the remaining attributes are considered as sensitive attributes. For example marital status and income are kept as sensitive attributes.

The Figure 1 shows the execution time of both algorithms and the Figure 2 shows the information losses of both algorithms. The Agglomerative hierarchical algorithm combined with t-closeness measure causes slightly less information loss and fast performance than the sequential clustering with the l-diversity measure.

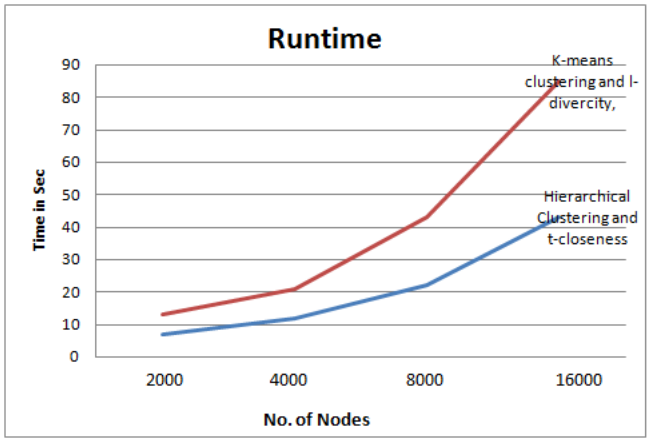

Figure 1: Runtime

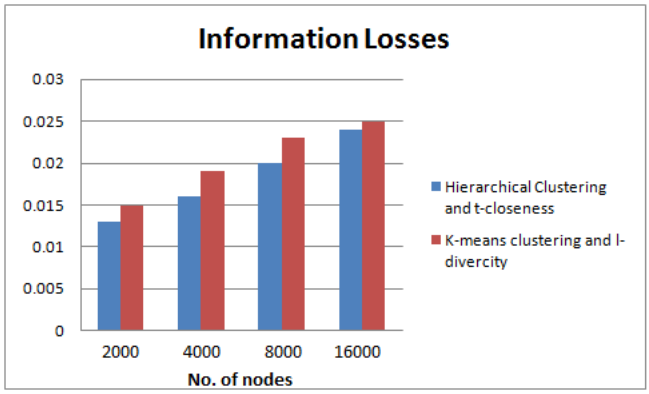

Figure 2: Information Losses

VIII. CONCLUSION

In this paper, the proposed a new anonymization approach for social network data to preserve privacy. We used some measures to find the information losses and developed the t-closeness privacy measure in combination with Agglomerative hierarchical clustering algorithm which outperforms the l-diversity privacy measure in combination with sequential clustering.

One research direction that this study suggests is to device
distributed versions of our proposed algorithms which might
require different techniques than those used here. Another research direction that this study suggests is to device a measure that combines the EMD with KL distance to improve the performance of the t-closeness measure.

REFERENCES

[1] A. Campan and T. M. Truta. A clustering approach for data and structural anonymity in social networks. In *Proceedings of the 2nd ACM SIGKDD International Workshop on Privacy, Security, and Trust in KDD (PinKDD'08), in Conjunction with KDD'08*, Las Vegas, Nevada, USA, 2008.

[2] 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.

[3] L. C. Freeman, D. R. White, and A. K. Romney. *Research Methods in Social Network Analysis*. George Ma-son University Press, Fairfax, VA, 1989.

[4] Jiawei Han, Micheline Kamber, *Data Mining: Concepts and Techniques*,2nd edition, Morgan Kaufmann, 2006.

[5] A.Machanavajjhala, D.Kifer, J.Gehrke and M.Venkitasubramaniam,"l-Diversity: Privacy Beyond K-anonymity", ACM Transactions on Knowledge Discovery and Data.vol.1, no.1, article 3, 2007.

[6] B.C.M Fung, K.Wang and P.S.Yu. "Top-down specialization for information and privacy preservation". In international conference on Data Engineering, 2005.

[7] Zheleva, E.,Getoor, "L.: Preserving the privacy on sensitive relationships in graph data". In: ACM SIGKDD Workshop on Privacy, Security, and Trust in KDD (PinKDD), pp. 153- 171(2007).

[8] Y.Rubner, C.Tomasi and L.J.Guibas. "The earth mover's distance as a metric for image retrieval". Int.J.Comput. Vision, 40(2):99-121, 2000.

[9] N.Li and T.Li . "t-closeness: Privacy beyond k-anonymity and l-diversity". In International Conference on Data Engineering (ICDE), 2007.

[10] C.R.Givens and R.M.Shortt. "A class of Wasserstein metrics for probability distributions". Michigan Math J., 31:231- 240, 1984.

[11] J.W.Byun, A.Kamra, E.Bertino, and N.Li. "Efficient k-anonymization using clustering techniques," In International Conference on Database Systems for Advanced Applications (DASFAA).2007.

[12] Ghinita, G., Karras, P., Kalinis, P., Mamoulis, "N.: Fast Data Anonymization with Low Information Loss". In: Very Large Data Base Conference (VLDB), pp.758–769 (2007).

[13] P.Samarati. "Protecting respondoent's privacy in microdata release". IEEE Transactions on knowledge and Data Engineering, 13, 2001.

[14] L.Sweeney. "Achieving k-anonymity privacy protection using generalization and suppression". International Journal of Uncertainty, 10(5):571–588, 2002.

[15] K.LeFevre, D.DeWitt and R.Ramakrishnan. "Mondrain multi dimensional k-anonymity". In Proc. 22nd International Conference Data Engineering (ICDE), 2006.

