ISSN:0975-9646

Anonymization of Centralized and Distributed Social Networks by Incremental Clustering

Ms. Sonali M. Khairnar[1], Prof. Sanchika Bajpai[2]

[1]*Department of Computer Engineering*
JSPM's BSIOTR (W), Wagholi Pune, India
[2]*Assistant Professor, Department Of Computer Engineering*
JSPM's BSIOTR (W), Wagholi Pune, India

***Abstract*— The social media has grown very vastly in the earlier years known think for all. There are different social media sites like Facebook, Twitter, LinkedIn, Google+ and many more that holds public and confidential/ personal information about their users. It is mandate to provide security to those users. In social network graphs are anonymized before being published to the others might be third person, for data mining or statistical study of privacy preservation of social networks. The purpose is to reach at an anonymized sight of the network without any of the data holders' information near links among nodes that are organized by other data holders'. For that our work start with the centralized setting and offer two variants of an anonymization algorithm which is previously based on sequential clustering and now we use Incremental Clustering for better performance. This algorithm produces anonymizations by means of clustering with better utility than those achieved by existing algorithms. We devised a secure distributed version of our algorithms for the case in which the network data is split between several players. We focused on the scenario in which the interacting players know the identity of all nodes in the network, but need to protect the structural information (edges) of the network.**

***Keywords*— Privacy Preserving Data Mining, Social Network, Clustering etc..**

I. Introduction

A network is a structure a set of devices (often referred to as nodes) connected by communication links referred as edges. A node can be a computer, printer, or any other hardware capable of sending and/or receiving data generated by other nodes on the network. A link can be a cable, air, optical fiber, or any other medium which can transport signal carrying information. Basically networks are modeled by a graph, where the nodes of the graph correspond to the entities, while edges denote relations between them. Real social networks may be more complex or contain additional information. In this study, we deal with social networks where the nodes could be accompanied by descriptive data, and propose two novel anonymization methods of the third category (namely, by clustering the nodes). Our algorithms issue anonymized views of the graph with significantly smaller information losses than anonymizations issued by the algorithms. We also devise distributed versions of our algorithms and analyze their privacy and communication complexity.

We also devise distributed versions of our algorithms and analyze their privacy and communication complexity. Working on this paper started with formal definitions in and a survey of related work, we stay in the real of centralized networks and propose two variants of an anonymization algorithm which is based on sequential clustering and incremental clustering. Next to describe a distributed form of our algorithms that computes a k-anonymization of the unified network by invoking secure protocols. The results of our experiments are given in second last section that specifies result comparison in between sequential clustering and incremental clustering. Finally we concluded in last section that incremental is far better algorithm then sequential by outlining future research directions in the study of privacy preserving study.

GOAL & OBJECTIVES

In this project we have main aim is to present the extended method for Anonymization of Centralized and Distributed Social Networks by incremental Clustering with improved reliability and performance.

- To present the present new framework and methods.
- To present the practical simulation of proposed algorithms and evaluate its performances.
- To present the comparative analysis of existing and proposed algorithms in order to claim the efficiency.

The method is implemented for to present the extended method for Anonymization of Centralized and Distributed Social Networks by incremental Clustering with improved reliability and performance, to present new framework and method of incremental clustering. It presents the comparative analysis of existing system.

As the previous system have certain outcomes i.e. Lack of reliability and scalability, inefficient method for large network which is not adaptable.

Problem Definition

To find out the minimum loss of information and hide the private data from the other party efficiently. Several studies have pointed out weaknesses of the k anonymity model in the context of tabular data. The main weakness of k-anonymity is that it does not guarantee sufficient diversity in the private attribute in each equivalence class of indistinguishable records.

II. THEORAETICAL BASICS

Clustering is grouping of objects having similar attributes. Clustering can be considered the most important unsupervised learning problem; so, as every other problem of this kind, it deals with finding a structure in a collection of unlabeled data. A loose definition of clustering could be "the process of organizing objects into groups whose members are similar in some way". A cluster is therefore a collection of objects which are "similar" between them and are "dissimilar" to the objects belonging to other clusters. This can be shown with a simple graphical example:

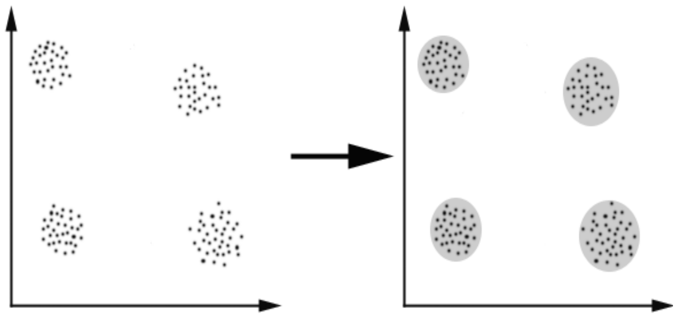

Fig. Graphical example of clustering

The algorithm is composed of the following steps:

1. Place K points into the space represented by the objects that are being clustered. These points represent initial group centroids.
2. Assign each object to the group that has the closest centroid.
3. When all objects have been assigned, recalculate the positions of the K centroids.
4. Repeat Steps 2 and 3 until the centroids no longer move. This produces a separation of the objects into groups from which the metric to be minimized can be calculated.

III. SYSTEM DESIGN AND IMPLEMENTATION

The presented algorithm applied on the dataset which forms the cluster find its centroid basis on the distance. Then it will go to initial portioning to find computation sum after checking loss of information.

Given a social network SN and a clustering C of its nodes, the information loss associated with replacing SN by the corresponding clustered network, SNC, is defined as a weighted sum of two metrics

$$I(C) = w \cdot ID(C) + (1 - w) \cdot IS(C);$$

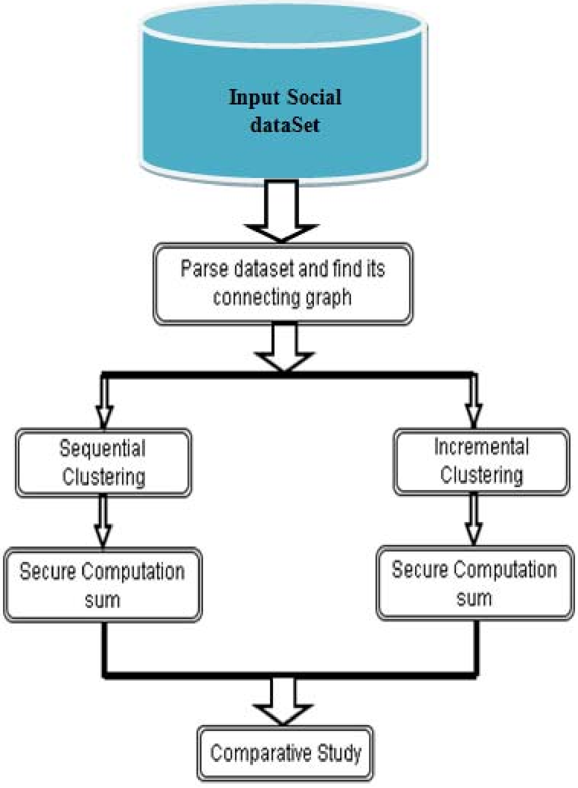

Fig. Architecture of the proposed system

ALGORITHM USED

1. Initial Partition Algorithm

Input: a SN Social network,an Integer k

Output: A clustering of SN into clusters of size ≥ k

Process:

```
step 1  Choose a random partition C={$C_1$,    ,$C_T$}of V into
   $T := \lfloor N / k_0 \rfloor$
Step  2. For n=1,....,N  do
Step 3. Let $C_t$ be the cluster to which $v_u$
currently belongs.
Step 4  For each of the other clusters, $C_s$, s≠t, compute
difference in  the information loss, $\Delta_{u:t\to s}$ if $v_u$ would
          move from $C_t$ to $C_s$.
Step 5.Let $C_{s0}$ be the cluster for which $\Delta_{n:t\to s}$ ,is minimal.
Step 6.If $C_t$is a singleton, move $v_n$ from $C_t$ to $C_{s0}$ and
          remove cluster $C_t$
Step 7.Else, if $\Delta_{n:t\to s0}$<0, movev$_n$ from $C_t$ to $C_{s0}$
Step 8.If there exist clusters cf size greater than $k_1$ split each
 of them randomly into two equally-sized clusters.
Step 9.If at least one node was moved during
the last loop,  go to Step 2 to 8.
```

Step 10.while there exist cluster of size smaller than k, select one of them and unify it with the cluster which is closest.
Step 11.resulting cluster as output

2: Singletone Algorithm

Input: Each player m, $1 \leq m \leq M$, has a private input vector $a_m \in Z_p^d$

Output: $a = \sum_{m=1}^{M} a_m$

Process:

Step 1.Player m select M random share vectors $a_{m,l} \in Z_p^d$ $1 \leq l \leq M$, such that $\sum_{l=1}^{M} a_{m,l} = a_m \bmod p.$

Step 2.Player m sends $a_{m,l}$ to the lth player, for all $1 \leq l \neq m \leq M$.

Step 3.Player l, $1 \leq l \leq M$, computes $s_l = \sum_{m=1}^{M} a_{m,l} \bmod p$

Step 4.Players l, $2 \leq l \leq M$, send s_l to player 1

Step 5.Player 1 computes $a = \sum_{l=1}^{M} s_l \bmod p$ and broadcast it.

3: Proposed Algorithm: Incremental Clustering algorithm

Input: Set of nodes (N), Number of clusters (k), Threshold

Output: A Clustering of SN

```
    1.  $cluster \leftarrow \emptyset$
    2.  For all $x_i \in N$ do
AS_F= False
For all $Clusters \in Clusters$ do
If || $X_i$ –centroid ( Cluster ) || < threshold  then
{
Update centroid (Cluster)
$X_i$
Ins_counter ( Cluster ) ++
AS_F=True
}
Exit loop
End if
End for
If (not AS_F)
{
Centroid ( newCluster ) =$X_i$
Ins_counter ( newCluster =1 )
$Cluster \leftarrow Cluster \cup newCluster$
}
End if.
    3.  End for
```

IV Experimental Work

Firstly we select input dataset that is one of XML file that contain information regarding users that is nodes.

Then we form the edges in form of graph by using the given input dataset which we select in previously. Later we take input to divide dataset into number of clusters. Shows Clusters are form by taking number of cluster size which is taken from user in previous state, along with that the edge instances is also form basis on nodes and edges.

After forming the cluster we go for next phase that is existing system algorithm which is sequential clustering, in that we perform partition and find computation sum. Initial partition is done at this level in that the partition forms as per the size of clusters formed.

Next basis on the partition we are going to find the single tone edges and then find which partition having minimum number of edges those edges to be moved, with the moving of edges it again finding loss of information of that edge. Then we shows updated clusters with updating of centroid, by taking new clusters as input value to find out computation sum of sequential clustering. With same we have concluded that value of "a" that refers to computation sum of sequential clustering. Again for the comparative study we take new input value for size of cluster to perform the same work as sequential clustering and finally find value of a that refers computation sum.

Then we shows updated clusters with updating of centroid, by taking new clusters as input value to find out computation sum of incremental clustering. With same we have concluded that value of "a" that refers to computation sum of sequential clustering. And the final working of the proposed system which shows the time required to the incremental clustering is less than the execution time required to incremental clustering.

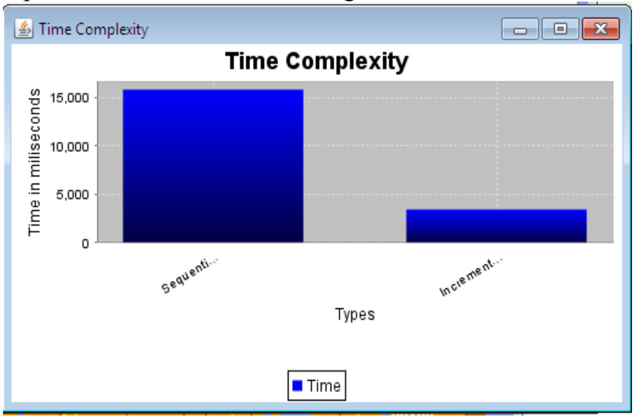

Fig: Resultant graph of Time complexity

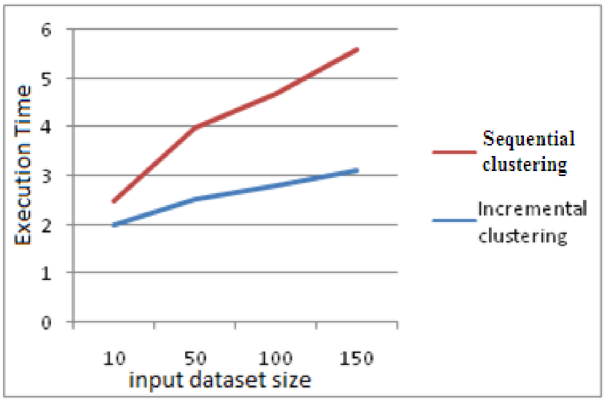

Fig: Resultant graph of Performance

The figure shows resultant graph of time complexity and performance of system. It first graph shows the required to execute our system less than the time required to existing system.

V. Conclusion

The novel real time model is contribution to existing historical model, where it presents the extended method for Anonymization of Centralized and Distributed Social Networks by incremental Clustering with improved reliability and performance.

The utility of our approach was demonstrated by running experiments on real and synthetic data set which improves the performance and reliability of the system. Our research suggests methods for quickly collecting information from the neighbourhood of a user in a dynamic social network when knowledge of its structure is limited or not available. We presented incremental clustering algorithms for anonymizing social networks. Those algorithms produce Anonymization by means of clustering with better utility than those achieved by existing algorithms.

We can also add future work to this system that the utility of our approach demonstrated by running experiments on real but offline and synthetic data sets. As the privacy protection implies we can block the users/attacker that is offline dataset for some time of span who is trying to hack the data on large dataset or online clustering.

VI. Acknowledgment

I wish to express my sincere thanks and deep gratitude towards my guide Prof. Sanchika Bajpai for her guidance, valuable suggestions and constant encouragement in all phases. I am highly 4 indebted to her help in solving my difficulties which came across whole Paper work. Finally I extend my sincere thanks to respected Head of the department Prof. G.M.Bhandari and Prof. A.C. Lomte [P.G Co-Ordinator] and all the people for their kind support and encouragement for this paper. Last but not the least, I wish to thank my Mother and fiancé for their unconditional love and support.

References

[1] 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, February 2013

[2] G. Aggarwal, T. Feder, K. Kenthapadi, R. Motwani, R. Panigrahy, D. Thomas, and A. Zhu, "Anonymizing Tables," Proc. 10th Int'l Conf Database Theory (ICDT), vol. 3363, pp. 246-258, 2005.

[3] L. Backstrom, C. Dwork, and J.M. Kleinberg, "Wherefore Art Thou r3579x?: Anonymized Social Networks, Hidden Patterns, and Structural Steganography," Proc. 16th Int'l Conf. World Wide Web (WWW), pp. 181-190, 2007.

[4] A. Baraba´si and R. Albert, "Emergence of Scaling in Random Networks," Science, vol. 286, pp. 509-512, 1999.

[5] J. Benaloh, "Secret Sharing Homomorphisms: Keeping Shares of a Secret Secret," Proc. Advances in Cryptology (Crypto), pp. 251-260, 1986.

[6] F. Bonchi, A. Gionis, and T. Tassa, "Identity Obfuscation in Graphs Through the Information Theoretic Lens," Proc. IEEE 27th Int'l Conf. Data Eng. (ICDE), pp. 924-935, 2011.

[7] A. Campan and T.M. Truta, "Data and Structural k-Anonymity in Social Networks," Proc. Second ACM SIGKDD Int'l Workshop Privacy, Security, and Trust in KDD (PinKDD), pp. 33-54, 2008.

[8] J. Goldberger and T. Tassa, "Efficient Anonymizations with Enhanced Utility," Trans. Data Privacy, vol. 3, pp. 149-175, 2010. [8] S. Hanhija¨rvi, G. Garriga, and K. Puolamaki, "Randomization Techniques for Graphs," Proc. Ninth SIAM Int'l Conf. Data Mining (SDM), pp. 780-791, 2009.

[9] M. Hay, G. Miklau, D. Jensen, D.F. Towsley, and P. Weis, "Resisting Structural Re-Identification in Anonymized Social Networks," Proc. VLDB Endowment (PVLDB), vol. 1, pp. 102-114, 2008.

[10] M. Hay, G. Miklau, D. Jensen, P. Weis, and S. Srivastava, "Anonymizing Social Networks," Univ. of Massachusetts, technical report, vol. 7, no. 19, 2007.

[11] V. Iyengar, "Transforming Data to Satisfy Privacy Constraints," Proc. Eighth ACM SIGKDD Int'l Conf. Knowledge Discovery and Data Mining (KDD), pp. 279-288, 2002.

[12] W. Jiang and C. Clifton, "A Secure Distributed Framework for Achieving k-Anonymity," The Int'l J. Very Large Data Bases, vol. 15, pp. 316-333, 2006.

[13] M. Kantarcioglu and C. Clifton, "Privacy-Preserving Distributed Mining of Association Rules on Horizontally Partitioned Data," IEEE Trans. Knowledge and Data Eng., vol. 16, no. 9, pp. 1026-1037, Sept. 2004.

[14] S. Kirkpatrick, D.G. Jr, and M.P. Vecchi, "Optimization by Simmulated Annealing," Science, vol. 220, no. 4598, pp. 671-680, 1983.

[15] K. Liu and E. Terzi, "Towards Identity Anonymization on Graphs," Proc. ACM SIGMOD Int'l Conf. Management of Data (SIGMOD), pp. 93-106, 2008.

[16] A. Machanavajjhala, D. Kifer, J. Gehrke, and M. Venkitasubramaniam, "'-Diversity: Privacy Beyond k-Anonymity," ACM Trans. Knowledge Discovery and Data, vol. 1, no. 1, article 3, 2007.

[17] M.E. Nergiz and C. Clifton, "Thoughts on k-Anonymization," Proc. Int'l Conf. Data Eng. (ICDE), p. 96, 2006.

[18] A. Schuster, R. Wolff, and B. Gilburd, "Privacy-Preserving Association Rule Mining in Large-Scale Distributed Systems," Proc. IEEE Int'l Symp. Cluster Computing and the Grid (CCGRID), pp. 411-418, 2004.

[19] N. Slonim, N. Friedman, and N. Tishby, "Unsupervised Document Classification Using Sequential Information Maximization," Proc. 25th Ann. Int'l ACM SIGIR Conf. Research and Development in Information Retrieval (SIGIR), pp. 129-136, 2002.

[20] L. Sweeney, "Uniqueness of Simple Demographics in the U.S. Population," Laboratory for Int'l Data Privacy (LIDAP-WP4), 2000.

Sonali M. Khairnar pursuing ME in Computer engineering, from University Of Pune India, received BE in Computer engineering from NMU Jalgaon India. Her research span is data mining, databases and network security.

Prof. Sanchika Bajpai received M.Tech from AMITY University, Lukhnow India, Joined JSPM's BSIOTR (W) Pune in 2012. Her research span is data mining, databases and information retrieval.