International Journal of Information & Computation Technology.
ISSN 0974-2239 Volume 4, Number 13 (2014), pp. 1315-1320

http://www. irphouse.com

Sequential Clustering for Anonymizing Social Networks

Santhi katari

Sree Vidyanikethan Engineering College
Tirupati Andhra Pradesh

ABSTRACT

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. Anonymization of the social networks (MySpace, Facebook, Twitter and Orkut) is essential to preserve privacy of information gathered by 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 sequential clustering algorithm for anonymizing a social network and a measure that quantifies the information loss in the anonymization process to preserve privacy. The algorithm significantly outperforms the SaNGreeA algorithm due to Campan and Truta which is the leading algorithm for achieving anonymity in networks by means of clustering. SaNGreeA builds the clustering greedily, one cluster at a time by selecting the seed node and then keep adding to it the next node. The main disadvantage of SaNGreeA is it does not contain any mechanism to correct bad clustering decisions which are made earlier and also it includes structural information loss which may be evaluated only when all of the clustering is defined. The sequential clustering algorithm does not suffer from those problems because in each stage of its execution it has a full clustering. It always makes decisions according to the real measure of information loss. Sequential clustering algorithm constantly allows the correction of previous clustering decisions.

Keywords Social networks, clustering, privacy preserving data mining, information loss.

1. 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 the links between them, which may describe relations of friendship, collaboration, correspondence, and so forth. Networks are modeled by a graph, where the node of the graph represents the entities and edges denote relationship between them. Real social networks are more complex or contain some additional information such as edges would be labeled and the graph nodes could be associated by attributes that provide demographic information i.e. age, gender, location or occupation etc.

Such social networks are of interest to researchers from many disciplines, be it sociology, psychology, market research, or epidemiology. However, the data in such social networks cannot be released as is, since it might contain sensitive information. Therefore, it is needed to anonymize the data prior to its publication in order to address the need to respect the privacy of the individuals whose sensitive information is included in the data. Data anonymization typically trades off with utility. Hence, it is required to find a golden path in which the released anonymized data still holds enough utility, on one hand, and preserves privacy to some accepted degree on the other hand.

In this paper we propose a novel anonymization technique based on clustering the nodes into super-nodes of size at least k, where k is the required anonymity parameter. The study of anonymizing social networks has concentrated so far on centralized networks, i.e., networks that are held by one data holder or player. But in distributed settings, the network data is split between several players.

This study deals with social networks where the nodes could be accompanied by descriptive data, and propose a novel anonymization method (namely, by clustering the nodes).that concerns anonymized views of the graph with significantly smaller information losses than anonymizations issued by the algorithms of [2] and [3].

II. Anonymization by clustering

In general we view the social network as a simple undirected graph, with G = (V, E), where V= $\{v_{1,}\ldots,v_N\}$is the set of nodes and E $\subseteq {}^{V}C_2$ is the set of edges. Each node represents to an individual in the graph and edge which connects two nodes describes the relationship between respective individuals. Each node in the social network graph is described by set of non identifying attributes like zip code or age, which are called quasi identifiers. Combination of such attributes could be used for unique identification by means of linking attacks [4]. Quasi identifiers are not themselves "unique Identifiers".

To summarize, a novel anonymized social network is defined as follows:

Definition 2.1. Let $A_1 . . . A_I$ be a collection of quasi identifier attributes. A social network over $V=\{v_{1,}\ldots,v_N\}$ is SN=(V,E,R) where E $\subseteq {}^{V}C_2$ is the structural data (edges),describing relationship between individuals in V, and $R=\{R_{1,\ldots,}R_N\}$, where $R_n \in A_1 X . . . X A_{I,} 1\leq n \leq N$, are the descriptive data of the individuals in V.

As in [1], [2], [3], we consider anonymizations of a given social network by means of clustering. Let $C = \{C_{1,}\ldots,C_T\}$ be a partition of V into disjoint subsets, or clusters; i.e., $V = Ut=1^T$ and $C_t \cap C_s =$ □ for all $1\leq t\neq s \leq T$. The corresponding

clustered graph $G_c = (V_c; E_c)$ is the graph in which the set of nodes is $V_c = C$, and an edge connects C_t and C_s in E_c iff E contains an edge from a node in C_t to a node in C_s. Each node $C_t \in V_c$ is accompanied by two pieces of information $|C_t|$ (the number of original V -nodes that C_t contains), and et, which is the number of edges in E that connect nodes within C_t. In addition, each edge $\{C_t; C_s\} \in E_c$ is labeled by a weight $e_{t,s}$ that stands for the number of edges in E that connect a node in C_t to a node in C_s.

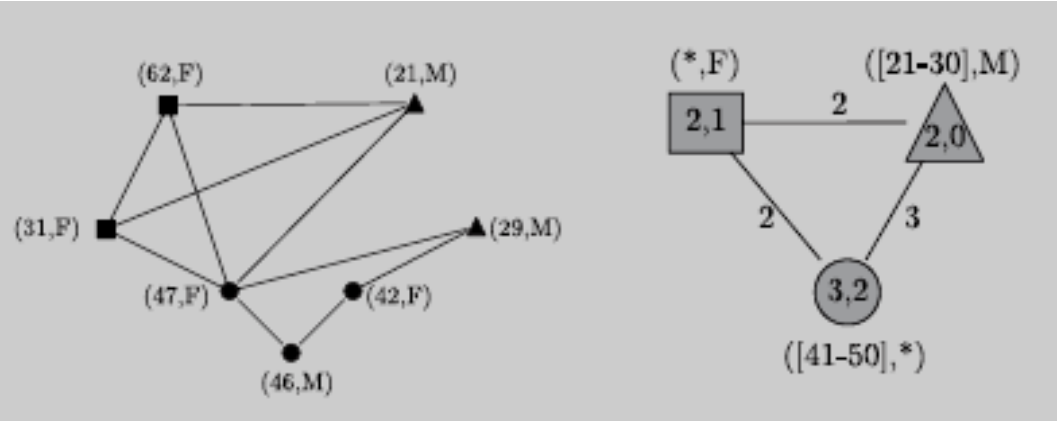

Fig1 : A network and a corresponding clustering

Let $G_c = (V_c, E_c)$ be a clustered graph that was derived from a graph G= (V ,E) of some social network SN =(V,E, R). Then, in addition to the structural data, which is given by E_c and the integral labels of the nodes, ($|C_t|$,e_t) and of the edges, $e_{t,s}$, one accompanies such a graph with descriptive data that is derived from the original descriptive data R. We apply the common method in anonymizing tabular data, and that is the generalization of the quasi-identifiers. Each of the quasi-identifiers, A_i, $1 \leq i \leq I$, is accompanied by a collection of subsets, A_i, which are the subsets of A_i that could be used for generalization.

Definition 2.2. Let SN= (V,E,R) be a social network and let $\bar{A}_{1,\ldots}$ $\bar{A}_I$ be generalization taxonomies for the quasi identifier attributes $A_1, \ldots, A_I$.Then given a clustering $C=\{C_{1,\ldots} C_T\}$ of V, the corresponding clustered social network is $SN=(C,V,E_c,R^I)$ where

$E_c \subseteq {}^{V}C_2$ is a set of edges on V_c , where $\{C_t ,C_s\} \in E_c$ iff there exist $V_n \in C_t$ and $V_n \in C_s$ such that $\{V_n, V_n\} \in E$;

The clusters in V_c are labeled by their size and the number of intra-cluster edges, while the edges in E_c are labeled by the corresponding number of inter-cluster edges in E;

$R^I=\{ R_1^I ,. . . , R_T^I \}$, where R^I_t is the minimal record in $\bar{A}_1 X_{\ldots} X \bar{A}_I$ that generalizes all quasi-identifier records of individuals in C_t, $1 \leq t \leq T$.

iii. SEQUENTIAL CLUSTERING ALGORITHM

The sequential clustering algorithm for k-anonymizing tables was presented in [7]. 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.

- Input: A social network $\mathcal{SN}$, an integer k.
- Output: A clustering of $\mathcal{SN}$ into clusters of size $\geq k$.

1) Choose a random partition $\mathcal{C} = \{C_1, \ldots, C_T\}$ of V into $T := \lfloor N/k_0 \rfloor$ clusters of sizes either k_0 or $k_0 + 1$.
2) For $n = 1, \ldots, N$ do:
   a) Let C_t be the cluster to which v_n currently belongs.
   b) For each of the other clusters, C_s, $s \neq t$, compute the difference in the information loss, $\Delta_{n:t\to s}$, if v_n would move from C_t to C_s.
   c) Let C_{s_0} be the cluster for which $\Delta_{n:t\to s}$ is minimal.
   d) If C_t is a singleton, move v_n from C_t to C_{s_0} and remove cluster C_t.
   e) Else, if $\Delta_{n:t\to s_0} < 0$, move v_n from C_t to C_{s_0}.
3) If there exist clusters of size greater than k_1, 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.

Fig 2: Sequential Clustering Algorithm

This algorithm starts with a random partition of the records into clusters. Then it goes over the n records in a cyclic manner and for each record checks whether it may be moved from its current cluster to another one while increasing the utility of the induced anonymization. This loop may be iterated when it reaches a local optimum (a stage in which no single record transition). As there is no guarantee that such procedure finds the global optimum, it may be repeated several times with different random partitions as the starting point in order to find the best local optimum among those repeated searches.

IV. A Modified Structural Information Loss Measure

Let B be the N X N adjacency matrix of the graph G=(V,E)i.e., $B(n,n^I)=1$ if $\{V_n,V_n^I\} \in E$ and $B(n,n^I)=0$ otherwise. Then, a Hamming-like distance is defined on V as follows:

$$D(n,n') := \frac{\left|\left\{\ell \neq n,n' : B(n,\ell) \neq B(n',\ell)\right\}\right|}{N-2}.$$

This definition of distance induces the following measure of structural information loss per cluster

$$I'_S(C_t) = \frac{1}{\binom{|C_t|}{2}} \cdot \sum_{v_n,v_{n'} \in C_t} \mathrm{D}(n,n') \ ,$$

The corresponding overall structural information loss is

$$I'_S(\mathcal{C}) = \frac{1}{N}\sum_{t=1}^{T} |C_t| \cdot I'_S(C_t) = \sum_{t=1}^{T} x(C_t),$$

Where

$$x(C_t) = \frac{2}{N(|C_t|-1)} \sum_{v_n,v_{n'} \in C_t} \mathrm{D}(n,n') \ .$$

In other words, I^I_S of a given cluster is the average distance between all pairs of nodes in that cluster, and I^I_S of the whole clustering is the corresponding weighted average of structural information losses over all clusters. For calculating descriptive information loss we use here LM metric. The LM metric associates the following loss of information with each of the nodes in that cluster,

$$I_D(C_t) = \frac{1}{I}\sum_{i-1}^{I} \frac{|R_i(i)|-1}{|A_i|-1};$$

The overall LM information loss is the result of averaging those losses of information over all nodes in V, i.e.

$$I_D(\mathcal{C}) = \frac{1}{N} \cdot \sum_{t=1}^{T} |C_t| \cdot I_D(C_t) \ .$$

Finally total weighted measure of information loss is then

$$I'(\mathcal{C}) = w \cdot I_D(\mathcal{C}) + (1-w) \cdot I'_S(\mathcal{C}) \ ,$$

Where w $\in$ [0,1]. Whenever the sequential clustering algorithm implements one of its decisions—be it moving a node from one cluster to another, splitting a large cluster, or unifying two small clusters—all that is needed in order to update I0 is to

update the intracluster information loss measures of the two clusters that are involved in such an action; there is no need to update also the intercluster information loss measures that involve all other clusters (as is the case when using I). This is why the number of cost function evaluations that sequential clustering needs to perform reduces from $O(N^3)$ to $O(N^2)$, when switching from I to I^I, in similarity to the SaNGreeA algorithm.

V. CONCLUSION

Sequential clustering algorithm is used for anonymizing 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. Sequential clustering algorithm produces anonymization by means of clustering with better utility than those achieved by existing algorithms. The main contributions in this paper are anonymizing social networks by using sequential clustering algorithm and a measure that quantifies the information loss in the anonymization process to preserve privacy.

References

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

[2] 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.

[3] E. Zheleva and L. Getoor, "Preserving the Privacy of Sensitive Relationship in Graph Data," Proc. ACM SIGKDD First Int'l Conf.Privacy, Security, and Trust in KDD (PinKDD), pp. 153-171, 2007.

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

[5] 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.

[6] 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.

[7] J. Goldberger and T. Tassa, "Efficient nonymizations with Enhanced Utility," Trans. Data Privacy, vol. 3, pp. 149-175, 2010.

[8] B. Zhou and J. Pei, "Preserving Privacy in Social Networks against Neighborhood Attacks," Proc. IEEE 24th Int'l Conf. Data Eng. (ICDE), pp. 506-515, 2008.