ISSN : 2347 - 8446 (Online)
ISSN : 2347 - 9817 (Print)
Vol. 2 Issue Special 1 Jan-March 2014
International Journal of Advanced Research in Computer Science & Technology (IJARCST 2014)

Preserving the Privacy on Social Networks by Clustering Based Anonymization

[I]J.Poulin, [II]M.Mathina Kani
[I]PG Student, [II]Assistant Professor
[I,II]Dept. of CSE, Sethu Institute of Tech., Affiliated to Anna University, Tamilnadu, India

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 informations 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 a SPKA 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, Clustering, Information Loss, Privacy Measure

I. Introduction

The Social networks are modelled as a graph which consists of nodes and edges. Nodes represent the set of individual's information such as age, gender, address. Edges represent the relationship between aforesaid individuals. The social network datasets are used by researchers from many disciplines. The field of sociology, psychology, market research, business analysis and epidemiology makes use of the social network informations [2]. In social networks the data sharing requires balancing many privacy and security. 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 SPKA 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. Clustering Based Anonymization

The primary goal in releasing the anonymized database for data mining is to deduce methods of predicting the private data from the public data [9]. This paper described the sequential clustering algorithm for k-anonymization. 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.

The social network data has begun to be analysed 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 [2]. This paper contributed the SaNGreeA (Social Network Greedy Anonymization) Algorithm which performs a greedy clustering processing to generate k-anonymous masked social network. This algorithm establishes good partitioning of all nodes from n into clusters. Next all nodes within each cluster are made uniform with respect to the quasi-identifier attributes. This homogenization is achieved by using generalization. Then this paper also introduces a measure to quantify generalization and loss of information.

The publishing data about individuals without revealing sensitive information about them is an important problem [3]. This paper proposed a novel and powerful privacy definition 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 atleast one well represented values 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. The SPKA algorithm

This paper proposes the SPKA (Single pass K-means Anonymization) Algorithm for k-anonymization. It derives from the K-means clustering algorithm but it involves single iteration instead of multiple iteration. Let D denotes the set of records to be anonymized and n be the number of records and K denotes the number of clusters to be generated.

The SPKA algorithm first sorts all the records in D by their quasi-identifiers. Then randomly select K records as the initial cluster centers to generate K clusters. For each record $r \in D$, it assigns r to the closest cluster center. Whenever a new record added to the cluster, the cluster centers get updated. Whereas in K-means

ISSN : 2347 - 8446 (Online)
ISSN : 2347 - 9817 (Print)

clustering algorithm after the complete cluster has formed, the cluster centers get updated. So that each cluster center reflects the current cluster center more accurately and consequently, the quality of following assignment of the records to the clusters get improved.

After clustering some adjustments are made on that to reduce the information losses. From the clusters with more than k records some records are removed and subsequently added to the clusters with less than k records. If all clusters contain no less than k records and still there are some records have not been assigned to any cluster means those records will be assigned to their respective closest clusters. The overall complexity of this clustering is $O(n^2/k)$.

SPKA ALGORITHM

Input : D Dataset: the value k for k-anonymity
Output : K clusters C = $\{C_1, C_2, \ldots C_k\}$ of D

1. Sort all records in D by their quasi-identifiers;
2. Randomly select K distinct records $r_1,\ldots,r_k \in D$;
3. Let $C_i := \{r_i\}$ for i=1 to K;
4. Let D :=D / $\{r_1,\ldots r_k\}$;
5. While (D≠0) do
   a. Let r be the first record in D;
   b. Calculate the distance between r to each C_i;
   c. Add r to its closest C_i;
   d. Update the center of the cluster;
   e. Let D:=D/{r};
6. End of While
7. For each cluster C with |C|>k do
   a. Sort all records in C;
8. While (|C| > k)
   a. Remove the record r€ C farthest from the centroid of C;
   b. Add it to the cluster where |C|<k
   c. Let C:=C/{r}; R:=R U {r};
9. End of While
10. End of For

IV. Generalization & Suppression

Generalization means replacing quasi-identifiers [2], the set of non identifying attributes such as age or zipcodes with less specific and less informative but semantically consistent values. For example the zipcodes {47824, 47865, 47843} can be generalized to 478** by stripping the rightmost digits. Various generalization strategies have been proposed in [4, 11, 12]. 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 [5].

Suppression does not release the value at all. 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. Those are Generalization information loss measure and Structural information loss measure. The first measure describes how much descriptive data details are lost during generalization. The second measure describes how much structural details are lost during generalization.

The generalization information loss measure was introduced in [1, 9, 10, 13]. In this paper, the generalization information loss measure is used as described in [1]. Let C_a 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

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

Where
N - Number of records in the table.
K_i - The size of each 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 C_a, $1 \le a \le T$, the original graph is replaced by the number of nodes K_a in the cluster C_a and number of edges E_a in the cluster C_a. Then the intra cluster information loss is measured as

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

Given two clusters C_a and C_b the structure of the edges that connects the nodes in C_a to the nodes in C_b 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

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

Where as shown in [2] SI(C) ranges between zero and one.

VI. t-closeness: A Privacy Measure

A novel privacy measure t-closeness is proposed in [7]. 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 [6]. It is a Monge-Kantorovich transportation distance [8] measure. 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 similarity attack which is the drawback of the l-diversity measure [3]. It protects against attribute disclosure. In this work we use both t-closeness and k-anonymity using SPKA at the same time for better results.

VII. Experimental Results

This section compares the performance of the proposed SPKA algorithm including t-closeness privacy measure against the combination of sequential clustering algorithm and l-diversity privacy measure. Both algorithms are implemented in C#.Net and run on a desktop PC with Intel coreTM i3 processor, 246MB of RAM under the Windows 7 Ultimate operating system.

This experiment uses the konet-sws dataset. Seven attributes of the taken dataset are treated as the quasi identifiers including age, gender, education, work class, zipcodes, race and country then the remaining attributes are considered as sensitive attributes. For example marital status and income are kept as sensitive attributes. Among these the attributes age, income, zipcodes are treated as numerical attributes, while the others are treated as categorical attributes.

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

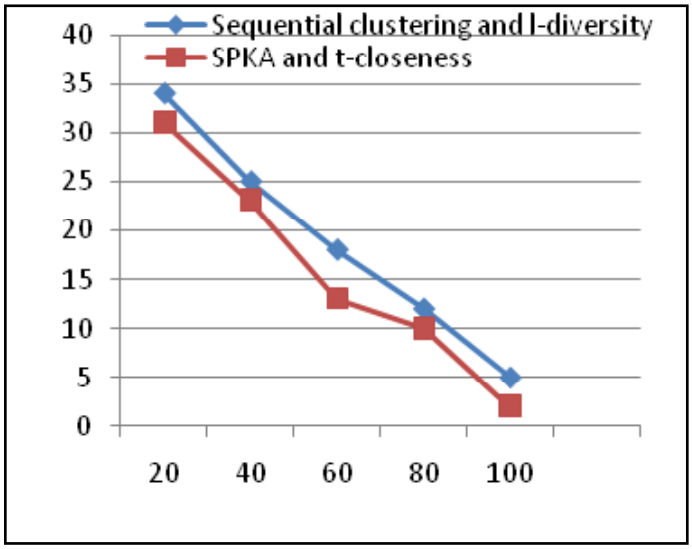

Figure 1: Runtime

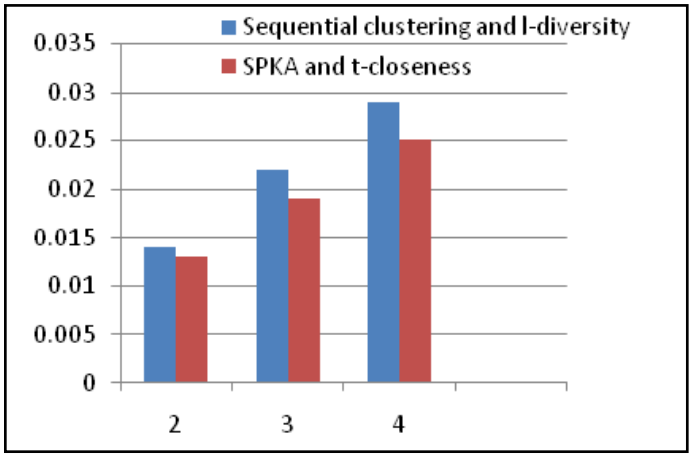

Figure 2: Information Losses

VIII. Conclusions and Future work

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 SPKA 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.

ISSN : 2347 - 8446 (Online)
ISSN : 2347 - 9817 (Print)

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, 2013.*

[2] *A.Campan and T.M.Truta,"Data and Strutural k-Anonymity in Social Networks," Proc.Second. ACM SIGKDD International Workshop on Privacy, Security, and Trust in KDD (PinKDD), pp.33-54, 2008.*

[3] *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.*

[4] *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.*

[5] *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).*

[6] *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.*

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

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

[9] *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.*

[10] *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).*

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

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

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

Poulin.J received her B.Tech degree in Information Technology from Sethu Institute of Technology, India, in 2012. She is currently pursuing her M.E in Computer science and Engineering from Sethu Institute of Technology, India. Her research interest includes Datamining.

MathinaKani.M received her B.E degree in Computer Science and Engineering from Dr.Sivanthi Aditanar College of Engineering, India, in 2001. She also received her M.E degree in Computer Science and Engineering from Arulmigu Meenakshi Amman College of Engineering, India, in 2005. Her research interest includes Datamining.