Closeness: An Advanced and Effective Measure for Data Publishing

Y Leela Sandhya Rani [1], M.Krishna[2], Shaik Nusrath Jahan[3]
[1]*Assistant Professor, Sir C. R. Reddy College of Engineering, Eluru, West Godavari Dt, AP, India*
[2]*Associate Professor, Sir C. R. Reddy College of Engineering, Eluru, West Godavari Dt, AP, India*
[3]*M. Tech Student, Sir C. R. Reddy College of Engineering, Eluru, West Godavari Dt, AP, India*

Abstract - The k-anonymity privacy requirement for publishing micro data requires that each equivalence class (i.e., a set of records that are indistinguishable from each other with respect to certain "identifying" attributes) contains at least k records. Recently, several authors have recognized that k-anonymity cannot prevent attribute disclosure. The notion of '-diversity has been proposed to address this; '-diversity requires that each equivalence class has at least 'well-represented (in Section 2) values for each sensitive attribute. In this paper, we show that '-diversity has a number of limitations. In particular, it is neither necessary nor sufficient to prevent attribute disclosure. Motivated by these limitations, we propose a new notion of privacy called "closeness." We first present the base model closeness, which requires that the distribution of a sensitive attribute in any equivalence class is close to the distribution of the attribute in the overall table (i.e., the distance between the two distributions should be no more than a threshold t). We then propose a more flexible privacy model called closeness that offers higher utility. We describe our desiderata for designing a distance measure between two probability distributions and present two distance measures. We discuss the rationale for using closeness as a privacy measure and illustrate its advantages through examples and experiments.

Keywords - *Closeness, Micro aggregate data, Incomplete data.*

I. INTRODUCTION

Data mining is a technique that deals with the extraction of hidden knowledge from large database. It uses sophisticated algorithms for the process of sorting through large amounts of data sets and picking out relevant information [1]. With the amount of data doubling each year, more data is gathered and data mining is becoming an increasingly important tool to transform this data into information [2]. The applications of data mining [4] includes wide range of areas as, credit card fraud detection, financial forecasting, automatic abstracting, medical diagnosis, analysis of organic compounds etc [6]. Data mining deals with large database which can contain sensitive information. An individual's private information is one of the example for sensitive information. It requires data preparation which can uncover information or patterns which may compromise confidentiality and privacy obligations [11]. Advancement of efficient data mining technique has increased the disclosure risks of sensitive data.

II. PROBLEM DEFINITION

One problem with l-diversity is that it is limited in its assumption of adversarial knowledge. As we shall explain below, it is possible for an adversary to gain information about a sensitive attribute as long as she has information about the global distribution of this attribute. This assumption generalizes the specific background and homogeneity attacks used to motivate diversity. Another problem with privacy-preserving methods, in general, is that they effectively

Assume all attributes to be categorical; the adversary either does or does not learn something sensitive. Of course, especially with numerical attributes, being close to the value is often good enough. In this project, we propose a novel privacy notion called "closeness." We first formalize the idea of global background knowledge and propose the base model t-closeness which requires that the distribution of a sensitive attribute in any equivalence class to be close to the distribution of the attribute in the overall table (i.e., the distance between the two distributions should be no more than a threshold t). This effectively limits the amount of individual-specific information an observer can learn. However, an analysis on data utility shows that t-closeness substantially limits the amount of useful information that can be extracted from the released data. Based on the analysis, we propose a more flexible privacy model called closeness, which requires that the distribution in any equivalence class is close to the distribution in a large-enough equivalence class (contains at least n records) with respect to the sensitive attribute. This limits the amount of sensitive information about individuals while preserves features and patterns about large groups. Our analysis shows that closeness achieves a better balance between privacy and utility than existing privacy models such as 'l-diversity and t-closeness. While the released table gives useful information to researchers, it presents disclosure risk to the individuals whose data are in the table. Therefore, our objective is to limit the disclosure risk to an acceptable level while maximizing the benefit. This is achieved by anonymizing the data before release. The first step of anonymization is to remove explicit identifiers. However, this is not enough, as an adversary may already know the quasi-identifier values of some individuals in the table. This knowledge can be either from personal knowledge (e.g., knowing a particular individual in person), or from other publicly available databases (e.g., a voter registration list) that include both explicit identifiers and quasi-identifiers. A common anonymization approach is generalization, which replaces quasi identifier values with values that are less-specific but semantically consistent.

As a result, more records will have the same set of quasi-identifier values. We define an equivalence class of an anonym zed table to be a set of records that have the same values for the quasi-identifiers. To effectively limit disclosure, we need to measure the disclosure risk of an anonym zed table. To this end, introduced k-anonymity as the property that each record is indistinguishable with at least k-1 other records with respect to the quasi-identifier.

In other words, k-anonymity requires that each equivalence class contains at least k records.

III. RELATED WORK

Same as for k-anonymity, the most common way to at tain t-closeness is to use generalization and suppression. In fact, the algorithms for k-anonymity based on those principles can be adapted to yield t-closeness by adding the t-closeness constraint in the search for a feasible minimal generalization: in the Incognito algorithm and in the Mondrian algorithms are respectively adapted to t-closeness. SABRE [2] is another interesting approach specifically designed for t-closeness. In SABRE the data set is first partitioned into a set of buckets and then the equivalence classes are generated by taking an appropriate number of records from each of the buckets. Both the buckets and the number of records from each bucket that are included in each equivalence class are selected with t-closeness in mind. One of the algorithms proposed in our paper uses a similar principle. However, the buckets in SABRE are generated in an iterative greedy manner which may yield more buckets than our algorithm (which analytically determines the minimal number of required buckets). A greater number of buckets leads to equivalence classes with more records and, thus, to more information loss. In an approach to attain t-closeness-like privacy isproposed which, unlike the methods based on generalization/suppression, is perturbative. Also, guarantees the threshold t only on average and uses a distance other than EMD. Another computational approach to t-closeness is presented in [8], which aims at connecting t-closeness and differential privacy; [8], also use a distance different from EMD but their method is non perturbative (the truthfulness of the data is preserved).Most of the approaches to attain t-closeness have been designed to preserve the truthfulness of the data. In this paper we evaluate the use of micro aggregation, a perturbative masking technique. In k-anonymity the relation between the quasi-identifiers and the confidential data is broken by making records in the anonymized data set indistinguishable in terms of quasi-identifiers within a group of k records. Micro aggregation, when performed on the projection on quasi-identifier attributes, produces a k-anonymous data set [9]. Micro aggregation was also used for k-anonymity without naming it in: clustering was used with the additional requirement that each cluster must have k or more records. While micro aggregation has been proposed to satisfy another refinement of k-anonymity (p-sensitive k-anonymity,), no attempt has been made to use it for t-closeness.

IV. BACKGROUND

Analyzing the three different approaches: k-Anonymity, t-closeness, micro aggregation. All the three approaches a basis for substantially reducing complexity by approximations.

k-Anonymity: An intruder re-identifies a record in an anonymized data set when he can determine the identity of the subject to whom the record corresponds. In case of re-identification, the intruder can associate the values of the confidential attributes in the re-identified record to the identity of the subject, thereby violating the subject's privacy.k-Anonymity seeks to limit the capability of the intruder to perform successful re-identifications. Definition 1 (k-anonymity). Let T be a data set and QI T be the set of quasi-identifier attributes in it. T is said to satisfy k-anonymity if, for each combination of values of the quasi-identifiers in QI T , at least k records in T share that combination. In a k-anonymous data set, no subject's identity can be linked (based on the quasi-identifiers) to less than k records. Hence, the probability of correct re-identification is, at most, 1/k. In what follows, we use the terms k-anonymous group or equivalence class to refer to a set of records that share the quasi-identifier values.

t-Closeness: Even though k-anonymity protects against identity disclosure, it is a well-known fact that k-anonymous data sets are vulnerable to attribute disclosure. Attribute disclosure occurs when the variability of a confidential attribute within an equivalence class is too low. In that case, being able to determine the equivalence class of a subject may reveal too much information about the confidential attribute value of that subject. Several refinements of k-anonymity have been proposed to deal with attribute disclosure. For example, p-sensitive k-anonymity , l-diversity, t-closeness, and (n,t)-closeness. As explained in Section 1, in this paper we focus on t-closeness because of its strict privacy guarantee (although the methods we propose are easily adaptable to (n,t)-closeness).t-Closeness seeks to limit the amount of information that an intruder can obtain about the confidential attribute of any specific subject. To this end, t-closeness requires the distribution of the confidential attributes within each of the equivalence classes to be similar to their distribution in the entire data set. Definition 2. An equivalence class is said to satisfy t-closeness if the distance between the distribution of the confidential attribute in this class and the distribution of the attribute in the whole data set is no more than a threshold t. A data set (usually a k-anonymous data set) is said to satisfy t-closeness if all equivalence classes in it satisfy t-closeness. The specific distance used between distributions is central to evaluate t-closeness, but the original definition does not advocate any specific distance. The Earth Mover's distance (EMD) is the most common choice (and the one we will adopt in this paper), although other distances have also been explored. EMD(P,Q) measures the cost of transforming one distribution P into another distribution Q by moving probability mass. EMD is computed as the minimum transportation cost from the bins of P to the bins of Q, so it depends on how much mass is moved and how far it is moved. For numerical attributes the distance between two bins is based on the number of bins between them. If the numerical attribute takes values {v 1 ,v 2 ,...,v m },where v i < v j if i < j, then ordered distance(v i ,v j) = |i − j|/(m − 1). Now, if P and Q are distributions over{v 1 ,v 2 ,...,v m } that, respectively, assign probability p iand q i to v i , then the EMD for the ordered distance can be computed as

$$EMD(P,Q) = \frac{1}{m-1}\sum_{i=1}^{m}\left|\sum_{j=1}^{i} p_j - q_j\right|$$

Microaggregation: Microaggregation is a family of perturbative methods for statistical disclosure control of microdata releases. One-dimensional microaggregation was introduced in [3] and multi-dimensional microaggregation was proposed and formalized in [5]. The latter is the one that is useful for k-anonymity and t-closeness. It consists of the following two steps: Partition: The records in the original data set are par-titioned into several clusters, each of them containing at least k records. To minimize the information loss, records in each cluster should be as similar as possible.

Aggregation: An aggregation operator is used tosummarize the data in each cluster and the original records are replaced by the aggregated output. For numerical data, one can use the mean as aggregation operator; for categorical data, one can resort to the median or some other average operator defined in terms of an ontology . The partition and aggregation steps produce some information loss. The goal of microaggregation is to minimize the information loss according to some metric. A common information loss metric is the SSE (sum of squared errors). When using SSE on numerical attributes, the mean is a sensible choice as the aggregation operator,because for any given partition it minimizes SSE in the aggregation step; the challenge thus is to come up with a partition that minimizes the overall SSE. Finding an optimal partition in multi-dimensional microaggregation is an NP-hard problem therefore, heuristics are employed to obtain an approximation with reasonable cost.The limitations to re-identification imposed by k-anonymity can be satisfied without aggregating the values of the quasi-identifier attributes within each equivalence class after the partition step. It is less utility damaging to break the relation between quasi-identifiers and confidential attributes while preserving the original values of the quasi-identifiers. This is the approach to attain k-anonymity-like guarantees taken in.

t-closeness through micro aggregation algorithm

Algorithm 1 consists of two clearly defined steps: first micro aggregate and then merge clusters until t-closeness is satisfied. In the micro aggregation step any standard microaggregation algorithm can be used because the enforcement of t-closeness takes place only after micro aggregation is complete. As a result, the algorithm is quite clear, but the utility of the anonymized data set may be far from optimal. If, instead of deferring the enforcement of t-closeness to the second step, we make the micro aggregation algorithm aware of the t-closeness constraints at the time of cluster formation, the size of the resulting clusters and also information loss can be expected to be smaller.

Algorithm 2 micro aggregates according to the above idea. It initially generates a cluster of size k based on the quasi-identifier attributes. Then the cluster is iteratively refined until t-closeness is satisfied. In the refinement, the algorithm checks whether t-closeness is satisfied and, if it is not, it selects the closest record not in the cluster based on the quasi-identifiers and swaps it with a recording the cluster selected so that the EMD to the distribution of the entire data set is minimized. Instead of replacing the records already added to a cluster, we could have opted for adding additional records until t-closeness is satisfied. This latter approach was discarded because it led to large clusters when the dependence between quasi-identifiers and confidential attributes is high. In this case, clusters homogeneous in terms of quasi-identifiers tend to be homogeneous in terms of confidential attributes, so the within-cluster distribution of the confidential attribute differs from its distribution in the entire data set unless the cluster is (nearly) as big as the entire data set. It may happen that the records in the data set are exhausted before t-closeness is satisfied.

k-Anonymity-first t-closeness aware micro aggregation algorithm.

```
function k-A NONYMITY - FIRST
Data: X: original data set
k: minimum cluster size
t: t-closeness level
Result Set of clusters satisfying k-anonymity and
t-closeness
Clusters = ∅
X 0 = X
while X 0 6= ∅ do
x a = average record of X 0
x 0 = most distant record from x a in X 0
C = GenerateCluster(x 0 , X 0 , X, k, t)
X 0 = X 0 \ C
Clusters = Clusters ∪ {C}
if X 0 6= ∅ then
x 1 = most distant record from x 0 in X 0
C = GenerateCluster(x 1 , X 0 , X, k, t)
X 0 = X 0 \ C
Clusters = Clusters ∪ {C}
end if
end while
return Clusters
end function
function G ENERATE C LUSTER (x, X 0 , X, k, t)
Data: x: source record for the cluster
X 0 : remaining unclustered records of X
X: original data set
k: minimum cluster size
t: desired t-closeness level
Result t-close cluster of k (or more) records
if |X 0 | < 2k then
C = X 0
else
C = k closest records to x in X 0 (including x
itself)
X 0 = X 0 \ C
while EMD(C,X) > t and X 0 6= ∅ do
```

```
y = record in X 0 that is closest to x
y 0 = record C that minimizes EMD(C∪{y}\
{y 0 },X)
if EMD(C ∪ {y} \ {y 0 },X) < EMD(C,X)
then
C=C ∪ {y} \ {y 0 }
end if
X 0 = X 0 \ {y}
end while
end if
return C
end function
```

t-Closeness Aware Micro Aggregation : t- Closeness - First

We modified the micro aggregation algorithm for it to build the clusters in a t-closeness aware manner. The clustering algorithm, however, kept the focus on the quasi-identifiers (records were selected based on the quasi-identifiers) and did not guarantee that every cluster satisfies t-closeness. The algorithm proposed in this section prioritizes the confidential attribute, thereby making it possible to guarantee that all clusters satisfy t-closeness. We assume in this section that the values of the confidential attribute(s) can be ranked, that is, be ordered in some way. For numerical or categorical ordinal attributes, ranking is straightforward. Even for categorical nominal attributes, the ranking assumption is less restrictive than it appears, because the same distance metrics that are used to micro aggregate this type of attributes can be used to rank them (e.g. the marginality distance in).We start by evaluating some of the properties of the EMD distance with respect to micro aggregation. To minimize EMD between the distributions of the confidential attribute within a cluster and in the entire data set, the values of the confidential attribute in the cluster must be as spread as possible over the entire data set. Consider the case of a cluster with k records. The following proposition gives a lower bound of EMD for such a cluster. Proposition 1. Let T be a data set with n records, A be a confidential attribute of T whose values can be ranked and C be a cluster of size k. The earth mover’s distance between C and T with respect to attribute A satisfies EMD A (C,T) ≥(n+k)(n−k)/(4n(n−1)k). If k divides n, this lower bound is tight. Proof. The EMD can intuitively be seen as the amount of work needed to transform the distribution of attribute A within C into the distribution of A over T. The “amount of work” includes two factors: (i) the amount of probability mass that needs to be moved and (ii) the n/k 2n/k 3n/k ... n (k-1)n/k c 1 c 2 c 3 c k. t-Closeness first, case k divides n. Confidential attribute values {c 1 ,c 2 ,...,c k } of the cluster C that minimizes the earth mover’s distance to T. When the confidential attribute values in T are grouped in k subsets of n/k values, c i is the median of the i-th subset for

i = 1,··· ,k. distance of the movement. When computing EMD for t-closeness, the distance of the movements of probability mass for numerical attributes is measured as the ordered distance, that is, the difference between the ranks of the values of A in T divided by n − 1.For the sake of simplicity, assume that k divides n. If that is not the case, the distance will be slightly greater, so the lower bound we compute is still valid. The probability mass of each of the values of A is constant and equal to 1/n in T, and it is constant and equal to1/k in C. This means that the first factor that determines the EMD (the amount of probability mass to be moved) is fixed. Therefore, to minimize EMD we must minimize the second factor (the distance by which the probability mass must be moved). Clearly, to minimize the distance, the i-th value of A in the cluster must lie in the middle of the i-th group of n/k records of T. Figure 1 illustrates this fact. In Figure 1 and using the ordered distance, the earth mover’s distance can be computed as k times the cost of distributing the probability mass of element c 1 among the n/k elements in the first subset:

$$\min(EMD) = k \times \sum_{i=1}^{n/k} \frac{1}{n} \frac{|i - {}^{n}/_{k}+{}^{1}/_{2}|}{n-1} = \frac{(n+k)(n-k)}{4n(n-1)k}$$

Above formula takes element (n/k + 1)/2 as the middle element of a cluster with n/k elements. Strictly speaking, this is only possible when n/k is odd. When n/k is even, we ought to take either b(n/k + 1)/2c, the element just before the middle, or d(n/k+1)/2e, the element just after the middle. In any case, the EMD ends up being the same as the one obtained in Formula (1).Note that, once n and t are fixed, Proposition 1 determines the minimum value of k required for EMD to be smaller than t. An issue with the construction of the k values c 1 , ···,c k depicted in Figure 1 is that it is too restrictive. For instance, for given values of n and t, if the minimal EMD value computed in Proposition 1 is exactly equal to t,then only clusters having as confidential attribute values c 1 , ···, c k satisfy t-closeness (there may be only one such cluster). Any other cluster having different confidential attribute values does not satisfy t-closeness. Moreover, in the construction of Figure 1, the clusters are generated based only on the values of the confidential attribute

V. SURVEY

Government agencies and other organizations often need to publish microdata, e.g., medical data or census data, for research and other purposes. Typically, such data are stored in a table, and each record (row) corresponds to one individual. Each record has a number of attributes, which can be divided into the following three categories:Attributes that clearly identify individuals. These are known as explicit identifiers and include, e.g., Social Security Number. Attributes whose values when taken together can potentially identify an individual. These are known as quasi-identifiers, and may include, e.g., Zip code, Birth-date, and Gender. Attributes that are considered sensitive, such as Disease and Salary. When releasing micro data, it is necessary to prevent the sensitive information of the individuals from being disclosed.

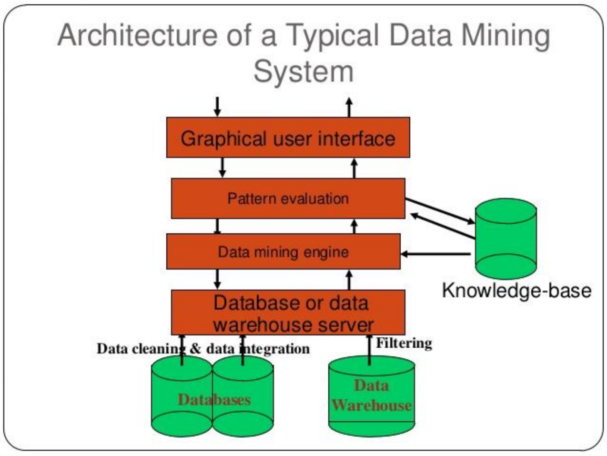
Architecture of a Typical Data Mining System
Graphical user interface
Pattern evaluation
Data mining engine
Database or data warehouse server
Knowledge-base
Data cleaning & data integration
Filtering
Databases
Data Warehouse

Fig-1: System Architecture

Security:

The protection k-anonymity provides is simple and easy to understand. If a table satisfies k-anonymity for some value k, then anyone who knows only the quasi-identifier values of one individual cannot identify the record corresponding to that individual with confidence greater than 1=k. While k-anonymity protects against identity disclosure, it does not provide sufficient protection against attribute disclosure.

Table1

	ZIP Code	Age	Disease
1	47677	29	Heart Disease
2	47602	22	Heart Disease
3	47678	27	Heart Disease
4	47905	43	Flu
5	47909	52	Heart Disease
6	47906	47	Cancer
7	47605	30	Heart Disease
8	47673	36	Cancer
9	47607	32	Cancer

Table2

	ZIP Code	Age	Disease
1	476**	2*	Heart Disease
2	476**	2*	Heart Disease
3	476**	2*	Heart Disease
4	4790*	≥ 40	Flu
5	4790*	≥ 40	Heart Disease
6	4790*	≥ 40	Cancer
7	476**	3*	Heart Disease
8	476**	3*	Cancer
9	476**	3*	Cancer

Example 1: Table 1 is the original data table, and Table 2 is an anonymized version of it satisfying 3-anonymity. The Disease attribute is sensitive. Suppose Alice knows that Bob is a 27-year old man living in ZIP 47678 and Bob's record is in the table. From Table 2, Alice can conclude that Bob corresponds to one of the first three records, and thus, must have heart disease. This is the homogeneity attack. For an example of the background knowledge attack, suppose that by knowing Carl's age and zip code, Alice can conclude that Carl corresponds to a record in the last equivalence class in Table 2. Furthermore, suppose that Alice knows that Carl has a very low risk for heart disease. This background knowledge enables Alice to conclude that Carl most likely has cancer.

Results

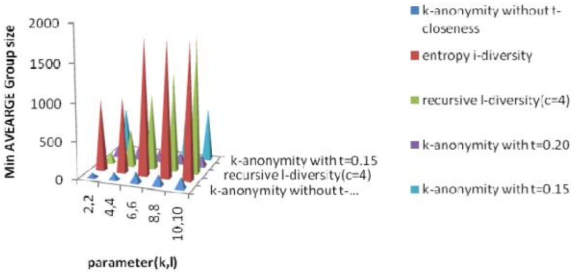

Fig-2: Average group size vs parameter

Privacy measure

In this project, we propose a novel privacy notion called "closeness." We first formalize the idea of global background knowledge and propose the base model t-closeness which requires that the distribution of a sensitive attribute in any equivalence class to be close to the distribution of the attribute in the overall table. This effectively limits the amount of individual-specific information an observer can learn. However, an analysis on data utility shows that t-closeness substantially limits the amount of useful information that can be extracted from the released data. This limits the amount of sensitive information about individuals while preserves features and patterns about large groups. To incorporate distances between values of sensitive attributes, we use the Earth Mover Distance metric to measure the distance between the two distributions. We also show that EMD has its limitations and describe our desiderata for designing the distance measure. We then propose a novel distance measure that satisfies all the requirements. Finally, we evaluate the effectiveness of the closeness model in both privacy protection and utility preservation through experiments on a real data set.

Data publishing

Privacy-preserving data publishing has been extensively studied in several other aspects. First, background knowledge presents additional challenges in defining privacy requirements. Second, several work considered continual data publishing, i.e., republication of the data after it has been updated. Presence to prevent membership disclosure, which is different from identity/attribute disclosure. Showed that knowledge of the anonymization algorithm for data publishing can leak extra sensitive information.

VI. CONCLUSION

Huge amount of data is collected everyday by many organization and individuals. The collected data are mined for knowledge discovery using numerous data mining algorithms. This raises serious concerns about privacy issues. A framework is developed for privacy preserving data mining which features high performance and strict privacy preserving algorithms.

VII. REFERENCES

[1]. L. Sweeney, Privacy Preserving Bio-Terrorism Surveillance, AAAI Spring Symposium, AI Technologies for Homeland Security, 2005.

[2]. E. Bertino, I. Fovino and L. Provenza, A Framework for Evaluating Privacy Preserving Data Mining Algorithms, Journal of Data Mining and Knowledge Discovery, 11(2), 2005, pp. 121–154.

[3]. M. Prakash, and G. Singaravel, A New Model for Privacy Preserving Sensitive Data Mining, Proceedings of Third International Conference on Computing Communication and Networking Technologies (ICCCNT), IEEE, 2012.

[4]. S. Verykios, E. Bertino, I. Fovino, L. Provenza, Y. Saygin and Y. Theodoridis, State-of-the-art in Privacy Preserving Data Mining, ACM SIGMOD Record, 33(1), 2004, pp. 50–57.

[5]. B. Pinkas, Cryptographic Techniques for Privacy Preserving Data Mining, ACM SIGKDD Explorations, 2002.

[6]. M. Prakash, and G. Singaravel, A Review on Approaches, Techniques and Research Challenges in Privacy Preserving Data Mining, Australian Journal of Basic and Applied Sciences, 8(10), 2014, pp. 251-259.

[7]. E. G. Komishani, and M. Abadi, A Generalization-Based Approach for Personalized Privacy Preservation in Trajectory Data Publishing, Proceedings of Sixth International Symposium on Telecommunications (IST'2012), IEEE, 2012.

[8]. R. Agrawal, and R. Srikant, Privacy Preserving Data Mining, ACM SIGMOD Conference, 2000.

[9]. Tiancheng Li and Ninghui Li, Slicing - A New Approach for Privacy Preserving Data Publishing, IEEE Transactions on Knowledge and Data Engineering, 24(3), 2012, pp. 561-574.

[10]. L. Sweeney, k-Anonymity: A Model for Protecting Privacy, International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems, 10(5), 2002, pp. 557-570.

[11]. M. Prakash, and G. Singaravel, An approach for prevention of privacy breach and information leakage in sensitive data mining, Journal of Computers and Electrical Engineering, In Press, DOIhttp://dx.doi.org/10.1016/j.compeleceng.2015.01.016, 2015.