# ADDITIVE DATA PERTURBATION APPROACH FOR PRIVACY PRESERVING DATA MINING

P. Chandrakanth[1], Dr. M. S. Anbarasi[2]
[1]Research Scholar, Pondicherry Engineering College, Puducherry
[2]Assistant Professor, Dept. of IT, Pondicherry Engineering College, Puducherry

**Abstract**

**In this big data era, large number of applications collects and analyzes personal data regularly. Sharing of these data is beneficial to the end users. Data is an important asset to business organizations and governments for decision making at the same time analyzing such data opens treats to privacy if not done properly. Privacy Preserving Data Mining aims to reveal the information by protecting sensitive data. Many methods such Randomization, k-anonymity and data hiding have been suggested in the erstwhile research. A systematic literature review has been done on the collection of articles and theses related to Privacy Preserving Data Mining and the quintessence of the survey is projected in this paper. Concepts of attacks, noise generation and perturbation methods have been discussed in the paper.**

**Index Terms: Privacy Preserving Data Mining, noise, K-Anonymity, Data perturbation, Additive Perturbation.**

## I. INTRODUCTION

Vast available digital data resources of information age leading to data collection and data mining demand a standard practice for those who aim; to efficiently discover relationships, patterns and association rules that are hidden in historical, varied formats and multiparty data; to predict future trends. These practices face challenges to their legal survival, how privacy of certain crucial data is preserved.

Data mining is a new area of research and a methodology that supports multidisciplinary technologies on data. Standard definition of data mining aims at finding valuable, qualitative, usable patterns of knowledge and information from large volumes of data. The current mixture of definitions, with each paper having its own definition of what "privacy" is maintained, will lead to confusion among potential adopters of the technology [9].

Much of commercial data analysis points on personal identity which directs towards customer centric business ideologies. Generalization involves replacing (or recoding) a value with a less specific but semantically consistent value. Suppression involves not releasing a value at all [7].

A social network is a universally wound graph structure; share most of the personal identifiable information in entities and connections between entities. The entities as nodes are the abstract representations of either individuals or organizations that are connected by links with one or more attributes. In high dimensional space the data becomes sparse, and the concept of spatial locality is no longer easy to define from an application point of view. In this paper, we view the k-anonymization problem from the perspective of inference attacks over all possible combinations of attributes. We show that when the data contains a large number of attributes which may be considered quasi-identifiers, it becomes difficult to anonymize the data without an unacceptably high amount of information loss [5].These methods shall also be provided, even diversely, to fit into large-scale computing environments and enable researchers and to create unified analysis framework.

### A. Individual Privacy

One of the key objectives of data privacy is the protection of personally identifiable information. Personally identifiable information is considered personally identifiable if it can be associated, directly or indirectly, to an individual

person. The attribute values associated with individuals are private and must be protected from disclosure, when attributes related personal data are subjected to mining. Data mining analysts allowed understanding a global models rather than from the characteristics of a particular individual.

B. Collective Privacy

Securing privacy for personal data is not just enough for a strategic privacy preserving model. Protect against learning sensitive knowledge representing the activities of a group is also essential. Collective privacy preservation is referred to as the protection of sensitive knowledge, achieving the goal for statistical databases by aggregating information and forming groups, furthermore prevent the disclosure of confidential information about individuals

## II. PRIVACY PRESERVING DATA PUBLISHING

The concept of privacy preserving data publishing is illustrated in the figure. The data publisher collects data from record owners in the data collection phase and releases to a data miner or to the public called data recipient in the data publishing phase. The data recipient will conduct data mining on the published data. The data recipient may extract inferences and evolve a data model for predictive techniques and patterns for descriptive techniques. From the observations and results of data publishing it is known that, two models are persistent, they are untrusted model and trusted model. In the untrusted model the data publisher is not trusted and may endeavor to identify sensitive information from record owners. Various cryptographic solutions, anonymous communication methods and statistical methods were proposed to collect records anonymously from the respective owners without jeopardizing the privacy of owner. In the trusted model, the publisher is believed to be trustworthy and record owners are willing to provide their personal information. The trust is not a transitive part to the data recipient. Therefore in this survey we assume the trusted model of data publishers and consider privacy issues in the data publishing. The issue of privacy preserving data mining Specifically, we consider a scenario in which two parties owning confidential databases wish to run a data mining algorithm on the union of their databases, without revealing any unnecessary information

That problem is a specific example of secure multi-party computation and, as such, can be solved using known generic protocols. Here focus on the problem of decision tree learning with the popular ID3 algorithm.[10]

## III. PRIVACY PRESERVING DATA MINING

Privacy Preserving Data Mining is a solution for data mining to increase sophistication of data mining algorithms to protect privacy. The research area, privacy-preserving data mining has become more important in recent years because of the increasing ability to store personal data about users, and the privacy preserving complexity in data mining algorithms. Techniques such as randomization, k-anonymity, etc., are also in vogue[2] since many years, in order to perform privacy-preserving data mining.

An optimal anonymization is one which perturbs the input dataset as little as is necessary to achieve -anonymity, where “as little as is necessary” is typically quantified by a given cost metric[6].

The balancing described above is premised on the ability to measure the probability of re-identification. Several metrics have been developed for measuring the probability of re-identification [1]. Privacy preserving data mining has the potential to increase the reach and benefits of data mining technology [9].

Privacy Preserving Data Mining Techniques

Techniques in Privacy Preserving Data Mining (PPDM) are primarily classified into three approaches, viz., perturbation, anonymization and cryptographic methods.

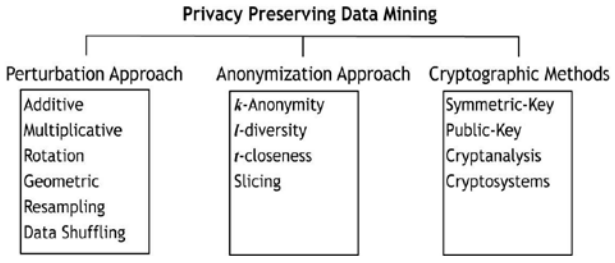


*Fig 1. Privacy Preserving Data Mining Technique*

Apart from the above classification as activities of PPDM are considered as post data mining and pre data mining activities, further the

methods for applying PPDM Randomization, Secure multi-party computation, sequential pattern hiding, data swapping, suppression, aggregation etc,.

## IV. RESEARCH DIRECTIONS IN PPDM

Research in PPDM is at tender stage, where with the advent of latest technologies of huge and high dimension data handling mechanism, the methodologies for applications of PPDM algorithms vary according to the data size, as obviously known the algorithmic complexities. To compete with this challenge many parallel algorithms also have been in the thought of research.

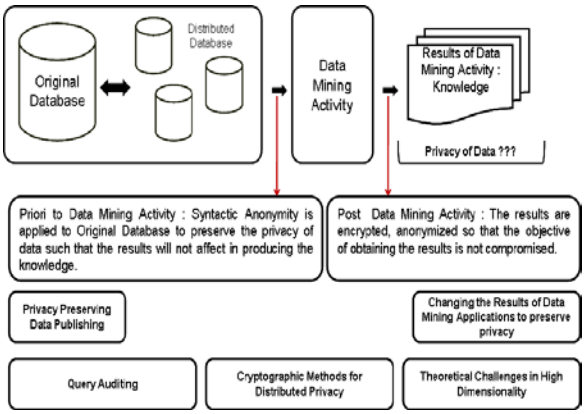


*Fig 2. Research directions in PPDM*

## V. PERTURBATION

Perturbation is a mechanism introduced in celestial mechanics and mathematics. a weight is associated with each attribute denoting its accuracy and completeness. Every constraint involving this attribute is associated to a weight representing the relevance of its violation[8].The more trusted a data miner is, the less perturbed copy of the data it can access. Under this setting, a malicious data miner may have access to differently perturbed copies of the same data through various means, and may combine these diverse copies to jointly infer additional information about the original data that the data owner does not intend to release. Preventing such diversity attacks is the key challenge of providing MLT-PPDM services[3]. novel reconstruction procedure to accurately estimate the distribution of original data values. By using these reconstructed distributions, we are able to build classifiers whose accuracy is comparable to the accuracy of classifiers built with the original data [4].

.

Thus perturbation mechanisms are ideally chosen for preserving privacy compared with other techniques. According to the anthological collections of Charu. C. Agrawal, et. al. “Privacy Preserving Data Mining : Models, Algorithms and Techniques”,[2] the perturbation methods are classified as: additive perturbation, multiplicative perturbation, rotation perturbation, geometric perturbation.

### *A. Perturbation Techniques*

As data perturbation is one common and unique approach that preserves privacy in data mining, the cost of implementation is very low and there is also an opening to the attackers. The actual research challenge lies in mitigating with the low cost (in terms of time and space) perturbation with more secured versions of data publications.

As a case study, Let us consider n ´ m is the original dataset, a real-valued matrix X containing columns and records. The data owner perturbs X to produce n′ ´ m data matrix Y, which is then release to the public or another party for analysis. The attacker uses Y and any other available information to produce an estimation of X, denoted by Xo.

1)Additive Perturbation: The data owner replaces the original data X with Y = X + R. Where R is a noise matrix with each column generated independently from a n-dimensional random vector r with mean vector zero. The entries in r were generated independently from some distribution with mean zero and significance variance with Gaussian distribution, where in this case r is referred to as additive white noise.

2)Multiplicative Perturbation: The data owner replaces the original data X with Y = MX. Where M is an n′ ´ n matrix chosen to have certain useful properties. If M is orthogonal, then the perturbation exactly preserves Euclidean distances, i.e., for any columns x1, x2 in X, their corresponding columns y1, y2 in Y. The advantage of this perturbation is that it is performed using a matrix, that it preserves Euclidean distance with either small or no error, it allows many important data mining algorithms to be applied to the perturbed data and produce results very similar to, or exactly the same as those produced by the original algorithm applied to the original data, e.g., hierarchical clustering,

k-means clustering. However, the issue of how well X is hidden is not clear and deserves a detailed study.

3)Rotation Perturbation: Rotation Perturbation is applied for Data Classification. This is a privacy model for all columns in the table. Rotation perturbation method considers the column privacy metric from the general privacy metric and an abstract privacy model is developed. Each column is also given by a privacy weight and based on the privacy weight the priority of perturbation is chosen.

4)Geometric Perturbation: Geometric data perturbation method is a linear combination, rotation, translation and distance based perturbation method. This method is comparatively used when alone rotation perturbation fails. Geometric perturbation in data-mining preserve the geometric class boundaries of the perturbed data.

## VI. TYPES OF ATTACKS

Deriving unofficially, illegally the personally identifiable information from the secure data bases is an attack. However, loss of personal identity of an individual secures the privacy, but the attackers and adversaries guess the probable values by joining the several perturbed copies and finds the most probable personal identifiers, thus forming an attack.

The attacks on the published data is classified into two types of prior knowledge:

known input-output: The attacker knows some small collection of original data records and the attacker knows the mapping between these known original data records and their perturbed counterparts in Y.

known-sample: the attacker has a collection of independent samples (columns of S) from X ( S may or may not overlap with X).

The privacy attacks are guessability attacks that are based on the known I/O prior knowledge assumption. The first one assumes an orthogonal perturbation matrix while the second assumes a randomly generated perturbation matrix. The third attack is based on the known sample prior knowledge assumption and assumes an orthogonal perturbation matrix.

### *A. Background knowledge attack:*

The structure or anatomy of attributes in a database releases all attributes to the user, where the end user or analyst must be able to pose the query. The attributes known from various queries and from the user access logs, the attackers can guess the structure of the database and administers an injection to extract data pertaining to the attributes. The group of attributes selected for perturbation or anonymization must be secure as a quasi-identifier table.

### *B. Minimality attack:*

The attacker combines two or more attributes to compose a quasi-identifier and tries to explore by random queries, there by extracting individuals' privacy data. In a quasi-identifier a group of related attributes may coexist, the attacker can derive the similar meaning from the names of the attributes and quasi-identifiers and draw information about the structure of data and attacks the original values in the database.

### *C. Unsorted Matching attack:*

Generally in a table that is to be published will have attributes perturbed in an unsorted order. The attacker guesses the values for the perturbed values in the database and arranges various combinations of sorting and finally can identify the relationship among the attribute values in the original data.

### *D. Temporal attack:*

Since data manipulation and changes to the data are dynamic, from the operations of add, delete and modify the attacker knows the nominal values of the attributes and guesses the structure of the data in the table.

### *E. Homogeneity attack:*

Attacker can guess the nature of other attributes with the knowledge of known attributes and thus finds the homogeneous attributes in the table.

## VII. TYPES OF NOISE

Noise in a dataset is an unstructured form. Like other series of data noisy data will also have a probabilistic distribution. Noise is generated synthetically using seed values just as random values. Noisy data generated using random values are always in normal distribution. The relevancy of noise with respect to the data set and domain meets the range of values and its entropy is controlled under criteria of minimum and maximum values of an attribute, such noise is fine grained noise. Noise that may relate to the

values of the attribute but with a maximum of entropy is coarse grained noise.

### *A. Gaussian Noise:*

Gaussian noise is statistical noise having a probability density function (PDF) equal to that of the normal distribution, which is also known as the Gaussian distribution. In other words, the values that the noise can take on are Gaussian-distributed.

### *B. Laplacian Noise:*

Laplacian Noise is statistically plotted with double exponential distribution. This distribution is also sometimes referred to Gumbel distribution. The difference of two competing noise series adding to a data base with a difference between two independent identically distributed exponential random variables is governed by a Laplace distribution.

Generally for deploying perturbation mechanism Gaussian and Laplacian noise are used. They have fully developed values of all ranges satisfying the standard deviation and mean, but there is no guarantee that they are fine grained noise. Since the values generated from Gaussian or Laplacian are function specific with domain seed values as input, they are coarse grained. Though Gaussian and Laplacian noise values have exponential precision, the type of values cannot be controlled in the Laplacian noise generation function, which is suitable only for real values.

Hence, selection of noise to deploy perturbation function in a data set shall not do major distortions in the data set, that the probability distribution of the original data set is not disturbed or modified. If the original data set and the perturbed data set contains similar distributions, it is very difficult for the adversary to guess the distribution of the noise from the perturbed data. Therefore, a smooth noise which can minimally change the values of the data set, such that their identity is not revealed shall be used for perturbations.

## VIII. GENERIC FRAMEWORK

A generic framework is proposed by the observation of all the current trends and available methodologies. The concept of privacy preserving data mining is in vogue with post and pre data mining stages. The presence of multi partition setup, multi-user group and the varied attackers and adversaries have thrown light into a resilient framework. The shown frame work describes a basic work flow in a development environment that consists of typical privacy preserving data mining activity. The data owner possesses data set that is more precious and useful for many stake holders, and data owner wants to distribute to the public use. The data owner may or may not be aware of the knowledge level of the public users, but should be aware that the data set published contains some personally identifiable information and needs privacy. Data owner observes some of the attributes as they belong to the privacy of individuals and allows perturbations. Selecting the attributes on the data set from the trusted server is the first phase of the experiment. The entire data base cannot be selected for perturbation; rather the number of attributes is selected based on the frequency of usage known from the user access log. The importance of the attributes used by the end user may vary timely based on the frequency of the usage of attributes also may change, so a classifier that can change its composition or the model criteria is selected that is called ensemble classifier. By using the machine learning and statistics, the information gain ratio is calculated for the attributes in the classifier and the level of usage or propensity of data in the algorithms that is how frequently such data will be projected to the public is calculated, based on this the sensitivity level of the data is estimated. The perturbation on the data set is performing based on the understanding of sensitivity levels of attributes in the dataset. Smooth noise with gradient properties may be used to perform perturbation. The following figure shows the framework and flow of activity in the privacy preserving data mining with syntactic anonymity and perturbation with smooth noise on sensitive data.

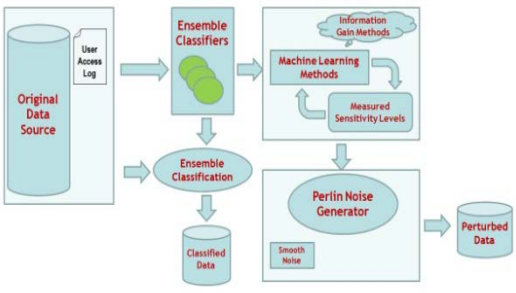


*Fig 3. Generic Framework of PPDM*

## IX. EXPERIMENTAL OBSERVATIONS

As a preliminary experimental setup in order to compare the anonymization and perturbation

we have generated hospital patient database synthetically. The database of patient information consists of patient id, patient name, age, gender, ailment, occupation. The anonymization and perturbation challenge is to give the data set to the analysts for estimating the ratio of gender base, age based descriptions for different ailments in their study. While disseminating this information to the analysts the database is perturbed as well as anonymized separately and the differences of the outcome are studied in the following graphs, in order to provide the good data quality for analyses.

### *A. Perturbation:*

The task of perturbation is not simple on the data sets. There must be a meticulous analysis for the selection of the portion of the data base and the selection of attributes to whom the perturbation shall be offered. Never, the entire data based is perturbed, except in the case of requirement of geometric perturbation with different sensitivity levels. A synthetic data set of 10000 records with 6 attributes has been used for the perturbation. For all the attributes compositional probability of their values is known. The min-max scaling method is applied on the selected attributes and normal values for each corresponding tuple is generated, where the source data can be plotted in the normal curve. The information gain methods are used on the attribute. A generic additive perturbation is compared with the perturbation framework state by us. For generic additive perturbation, randomly generated noise is applied, where in our framework Perlin noise is applied. Following graphs shows the performance of the experimentation.

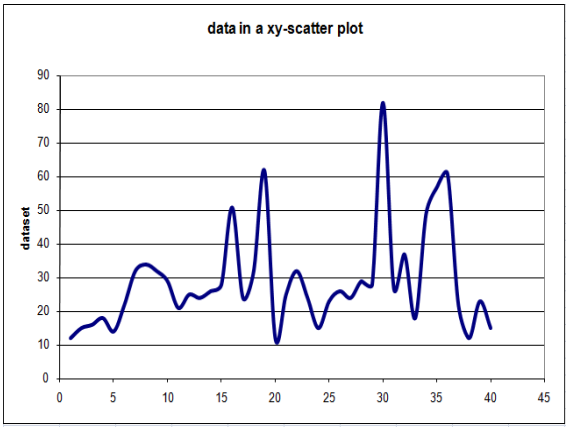


*Fig 4. Nature of data in the original source*

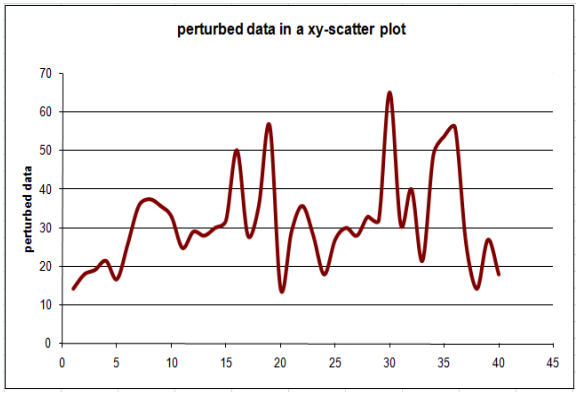


*Fig 5. Nature of data after perturbation using Additive Perturbation method*

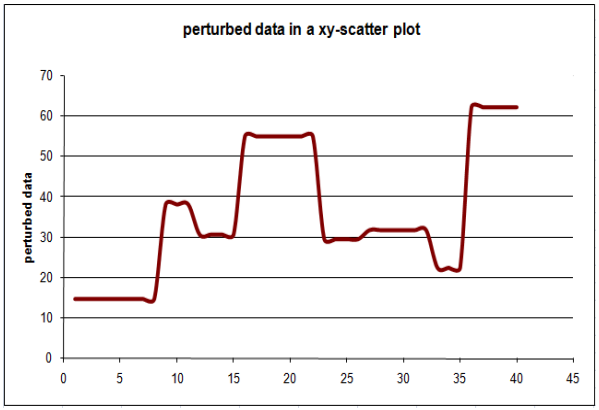


*Fig 6. Nature of data after perturbation using k-anonymization method*

### *B. Anonymization*

k-anonymity mode is developed to allow the trusted data user to perturb the data and disseminate. There are two methods of implementing the anonymity using suppression and generalization. Suppression mechanism uses the popular value in the model and replaces. The generalization uses the most frequent value to hide the other values of the datasets. One thousand simple random samples were drawn from each data set at nine different sampling fractions (0.1 to 0.9 in increments of 0.1). Any identifying variables were removed and each sample was k-anonymized.

### *C. Observations*

From the above graphs it is inferred that Additive perturbation methods are applied based on the scaled values of the data points, whereas in k-anonymity a model based suppression of values though the attempt is made for achieving good quality of data, but comparatively the additive perturbation methods have high and nearing quality of the original data sets.

The quality of the experimentation is proved good using the true-positive and false-negatives.

The analysts have achieved good results with the data sets perturbed using Additive Perturbation methods than the k-anonymity methods. However, the Additive perturbation methods do scan entire data sets, whereas the k-anonymity works only on the data that fits to the model.

$$precision = \frac{|\{total\ relevant\ data\}| \cap |\{total\ retrieved\ data\}|}{|\{total\ retrieved\ data\}|}$$

$$recall = \frac{|\{total\ relevant\ data\}| \cap |\{total\ retrieved\ data\}|}{|\{total\ relevant\ data\}|}$$

Precision is the average probability of relevant retrieval. Recall is the average probability of complete retrieval. Here we average over multiple retrieval queries.

The values after perturbation are analyzed from the methods as follows:

*Table 1. Precision-recall values calculated on k-anonymization algorithms*

| experimentation | precision | recall |
|---|---|---|
| 1 | 1000 | 543 |
| 2 | 1000 | 521 |
| 3 | 1000 | 478 |
| 4 | 1000 | 422 |
| 5 | 1000 | 380 |

*Table 2. precision-recall values calculated on additive perturbation algorithms*

| experimentation | precision | recall |
|---|---|---|
| 1 | 1000 | 990 |
| 2 | 1000 | 985 |
| 3 | 1000 | 987 |
| 4 | 1000 | 994 |
| 5 | 1000 | 993 |

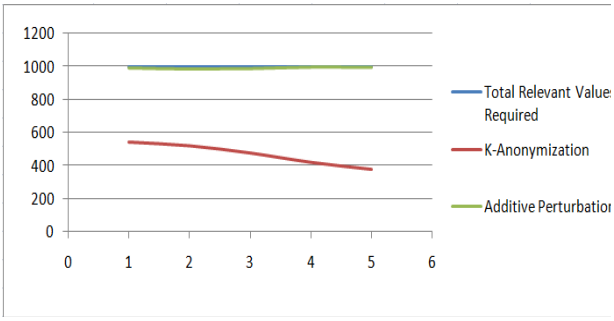


*Fig 7. A comparative graph of precision recall values of first few experimentations performed between k-anonymity and additive perturbation for a 1000 sample data set.*

## X. CONCLUSION

This survey is an outcome of a systematic literature review, that has been done on the collection of articles and theses related to Privacy Preserving Data Mining. Concepts of attacks, noise generation and perturbation methods have been discussed. Existing methods in terms of privacy models, anonymization operations, information metrics, and anonymization algorithms are compared. From survey we can draw research ideas and that concludes that the privacy preserving problem from different perspectives can help to decide appropriate privacy-preserving technology.

## REFERENCES

[1] El Emam K. Guide to the de-identification of personal health information. CRC Press, 2013.

[2] Charu C. Aggarwal and Philip S. Yu, “Privacy-Preserving Data Mining - Models and Algorithms”, © 2008 Springer Science+Business Media, LLC. ISBN: 978-0-387-70991-8 [524 pages].

[3] Yaping Li, Minghua Chen, Qiwei Li, and Wei Zhang, “Enabling Multilevel Trust in Privacy Preserving Data Mining”, IEEE Transactions On Knowledge And Data Engineering, Vol. 24, No. 9, Pp. 1598, © September 2012

[4] R. Agrawal, R. Srikant. “Privacy-preserving data mining”. In Proceedingsof the ACM SIGMOD Conference on Management of Data, pages439–450, Dallas, TX, May 2000.

[5] Charu C. Aggarwal. “On k-anonymity and the curse of dimensionality”. InProceedings of the 31st VLDB Conference, pages 901–909, Trondheim,Norway, 2005.

[6] Bayardo, R., Agrawal, R. “Data privacy through optimal k-anonymization”. In: Proc. of the21st International Conference on Data Engineering (2005).

[7] Sweeney, L.: “Achieving k-anonymity privacy protection using generalization

and suppression".International Journal of Uncertainty, Fuzziness and Knowledge Based Systems 10(5), 571–588 (2002).

[8] Chris Clifton and Murat Kantarcioglu and Jaideep Vaidya (2002), "Defining Privacy for Data Mining", Proceedings of the NationalScience Foundation Workshop on Next Generation Data Mining,pp.274-281.

[9] Igor Fovino and Marcelo Masera (2008), "Privacy Preserving Data Mining: A Data Quality Approach", JRC Scientific and TechnicalReports, Vol. 2, pp. 28-37.

[10] Y. Lindell, B. Pinkas, "Privacy Preserving Data Mining", Journal of Cryptology, vol. 15, no. 3, 2002, pp. 177-206