State-Of-The-Art in Refurbishing-Based Approach for Privacy Preserving Data Mining

J.Indumathi and G.V.Uma

Abstract—**Data mining, with its pledge to competently discern valuable, non-obvious information from bulky databases, is principally defenseless to misuse. In this paper we present an outline of the, new and rapidly emerging research area of Privacy Preserving Data Mining and the technical likelihood of realizing Privacy Preserving Data Mining. We also propose a classification hierarchy for refurbishing based techniques for privacy preservation that will lay down the origin for analyzing the toil which has been performed in this environment. A comprehensive appraisal of the toil accomplished in this vicinity is also given, along with the coordinates of each exertion to the classification hierarchy. For the purposes of this paper, we take for granted that suitable right of entry controls and security procedures are in position and effectual in preventing unconstitutional access to the system. Whenever sensitive information is exchanged, it must be transmitted in excess of a secure channel and stored securely. A short and snappy appraisal is performed, and some preliminary conclusions are made along with the metrics used for it.**

Index Terms— **Bayes Estimate, Bayes theorem, Correlation, Correlation Matrix, Data Mining, PCA, Privacy preserving, Privacy Preserving Data Mining, Randomization.**

INTRODUCTION

We now present a sketch out of the, new and rapidly emerging research area of privacy preserving data mining and the technical likelihood of realizing privacy preserving data mining. We also propose a classification hierarchy for reconstruction based techniques for privacy preservation that will lay down the origin for analyzing the toil which has been performed in this environment. A comprehensive appraisal of the sweat accomplished in this vicinity is also given, along with the coordinates of each exertion to the classification hierarchy. For the purposes of this work, we take for granted that suitable right of entry controls and security procedures are in position and effectual in preventing unconstitutional access to the system. Whenever sensitive information is exchanged, it must be transmitted in excess of a secure channel and stored securely. A short and snappy appraisal is performed, and some preliminary conclusions are made along with the metrics used for it.

Given the type of mystification added to the data, and the mystified data set, we must be able to reconstruct the original distribution (but not actual data values) of the data set. This facilitated a data mining algorithm to build a much more precise data than mining the mystified data alone.

Manuscript received December 31, 2007

Indumathi.J received her M.E. from Anna University, Chennai, India in year 1992 and M.B.A from Madurai Kamaraj University, Madurai, India in 1994. She is working for Anna University as a Senior Lecturer. (corresponding author to provide phone: 91-044-22432410; fax: 91-044-22432410; e-mail: indu@cs.annauniv.edu).

G.V.Uma, a Polymath received her M.E. from Bharathidasan University, India in year 1995 and Ph.D. from Anna University, Chennai, India in 2002. She is working for Anna University as an Assistant Professor. (e-mail: gvuma@annauniv.edu).

The crack of dawn of e gadgets like cell phones, laptops, palmtops, etc., is building extensive right to use to outsized magnitude of data probable. Sophisticated scrutiny of data for extracting functional information is the next normal stride in the planet of ubiquitously computing. Knowledge discovery and data mining contract with the problem of extracting interesting associations, classifiers, clusters, and other patterns from data. Data Mining is in concert a progressively more imperative role in business, scientific, and engineering applications because of the budding ease of use of data in electronic format. This will take on board Knowledge discovery into our steadily more mobile but attached way of life and proffer technology to scrutinize time-critical information collected at different locations from everywhere. This will pioneer a new cohort of applications in countless domains such as finance, health, service, manufacturing industries, and sensor networks for defence applications.

''Privacy Preserving Data Mining is getting valid data mining results without learning the underlying data values'', has been receiving attention in the research community and beyond. Privacy Preserving Data Mining has the potential to increase the reach and benefits of data mining technology. There has recently been a surge in interest in Privacy Preserving Data Mining [34, 7, 36, 25, 9, and 29]. Even the popular press has picked up on this trend [39, 1].

Utilizing this characteristic of ensuring that a data mining project won't facilitate *exploitation* of personal information opens opportunities that "complete privacy" would prevent. To do this, we call for technical and social solutions that guarantee data will not be on the rampage.

The rest of the paper is organized as follows. In Section 2, we first outline related literature survey and overview. In Section 3, we formally define the classification of privacy preserving techniques. In Section 4, we bring about a review of the privacy preserving techniques. In Section 5, we elaborately define the classification of refurbishing-based privacy preserving techniques. In Section 5, we discuss and classify the metrics for quantification of refurbishing-based privacy preserving techniques.

I. RELATED LITERATURE SURVEY AND OVERVIEW

There has been all-embracing delve into in the area of statistical databases stimulated by the yearning to be able to endow with statistical information (sum, count, average, maximum, minimum, p^{th} percentile, etc.) devoid of compromising susceptible information about persons. We can classify them based in the following ways:

1. Query Restriction and Data Perturbation
2. Mode of alteration of a data value by a new value
3. Methods adapted for modifying values of sensitive attributes

1. Query Restriction and Data Perturbation

The active techniques of query restriction and data perturbation may be classified into:

Query Restriction
Based on the restriction features of queries they can be further classified into:

- Restricting the size of query result (e.g. [19] [8])
- Controlling the overlap amongst successive queries (e.g. [21])
- Keeping audit trail of all answered queries and constantly checking for possible compromise (e.g. [37])
- Suppression of data cells of small size (e.g. [27])
- Clustering entities into mutually exclusive atomic populations (e.g. [17]).

2. Mode of Alteration of a data value by a new value
Data Perturbation
Based on the mode of alteration of a data value by a new value they can be classified into:

- Swapping values between records (e.g. [10])
- Replacing the original database by a sample from the same distribution (e.g. [18] [16] [15])
- Adding noise to the values in the database (e.g. [22] [11])
- Adding noise to the results of a query (e.g. [20])
- Sampling the result of a query (e.g. [31]).

3. Methods adapted for modifying values of sensitive attributes
Based on the methods for modifying values of sensitive attributes [9] we can classify into:

Value-Class Membership In this scheme, the values for an attribute is partitioned into a set of disjoint, mutually-exclusive classes. Consider the special case of Discretization in which values for an attribute are discretized into intervals. All intervals need not be of equal width. For example, rent may be discretized into 5K intervals for lower values and 50K intervals for higher values. As a substitute for of a true attribute value, the user provides the hiatus in which the value lies. Discretization is the means used for the most part of hiding individual values.

Value Distortion Return a value x_i + r instead of x_i where r is a random value drawn from some distribution. Consider two random distributions:

• **Uniform:** The random variable has a uniform distribution, between [-a, + a]. The mean of the random variable is 0.

• **Gaussian:** The random variable has a normal distribution, with mean p = 0 and standard deviation o".

There are negative results showing that the proposed techniques cannot convince the contradictory objectives of providing high quality statistics and at the same time prevent exact or partial disclosure of individual information [29].

II. PERSPECTIVES OF PRIVACY PRESERVING DATA MINING (PPDM) TECHNIQUES

There are many approaches which have been adopted for Privacy Preserving Data Mining . A prevalent classification is based on the perspectives of data distribution, data modification, data mining algorithm, data or rule hiding, privacy preservation. The last perspective which is the most imperative refers to the privacy preservation methods used for the discerning alteration of the data. Selective amendment is necessary in order to bring about eminent utility for the modified data known that the privacy is not given an impetus. The techniques that have been functional for this reason are:

• Modification-Based Techniques like adaptive alteration
• Cryptography-based techniques like secure multiparty computation
• Refurbishing-based techniques where the original distribution of the data is reconstructed from the randomized data.

It is vital to realize that data modification results in dilapidation of the database feat. In order to reckon the dilapidation of the data, we mainly use two metrics. The first one, measures the confidential data protection, while the second measures the loss of functionality.

III. REVIEW OF PRIVACY PRESERVING TECHNIQUES

A. Modification-Based Techniques

A number of techniques have been developed for a quantity of data mining techniques like classification, association rule discovery and clustering, based on the hypothesis that discerning data modification or sanitization is an NP-Hard problem, and for this basis, alteration can be used to address the complexity issues.

- Swapping values between records (e.g. [10])
- Replacing the original database by a sample from the same distribution (e.g. [18] [16] [15])
- Adding noise to the values in the database (e.g. [22] [11])
- Adding noise to the results of a query (e.g. [20])
- Sampling the result of a query (e.g. [31]).

B. Cryptography-based techniques

Another approach to achieve Privacy-Preserving Data Mining is to use Secure Multi-party Computation (SMC) techniques. Briefly, an SMC problem deals with computing certain function on multiple inputs, in a distributed network where each participant holds one of the inputs; SMC ensures that no more information is revealed to a participant in the computation than what can be inferred from the participant's input and the final output [13].Several SMC-based privacy-preserving data mining schemes have been proposed [6, 38 , 28, 40, 34]. Lindell and Pinkas used SMC to build decision trees over the horizontally partitioned data [6]. Vaidya and Clifton proposed the solutions to the clustering problem [40] and the association rule mining problem [28] for vertically partitioned data. Several SMC tools and fundamental techniques are also proposed in the literature [38]. Some more schemes were presented in recent conferences as follows. Wright et al. [26] and Meng et al. [2] used SMC to solve privacy-preserving Bayesian network problems. Gilburd et al. proposed a new privacy model, k-privacy, for real-world large-scale distributed systems [23]. Sanil et al. described a privacy-preserving algorithm of computing regression coefficients [33].. Wang et al. used an iterative bottom-up generalization to generate data, which remains useful to classification but difficult to disclose private sources [39].

C. Refurbishing-based techniques

Refurbishing-based techniques are techniques where the original circulation of the data is reconstructed from the randomized data.

Algorithm for Refurbishing

Step 1 : Creating randomized data replica by data perturbation of entity data records
Step 2 : Recreate distributions, not values in individual records.
Step 3 : By means of using the reconstructed distributions, fabricate the original data

For refurbishing use a fitting loom and suggest algorithms for edifice original data that rely on reconstructed distributions.

The predicament of reconstructing original distribution from a given distribution can be viewed in the broad-spectrum framework of inverse problems [30]. In [12], it was publicized that for smooth sufficient distributions (e.g. slowly varying time signals), it is possible to fully recuperate original distribution from non-overlapping, adjoining partial sums. Such partial sums of true values are not available to us. We cannot formulate priori assumptions about the original distribution; we merely be acquainted with the distribution used in randomizing values of an attribute. There is rich query optimization literature on estimating attribute distributions from partial information [4]. In the OLAP literature, there is work on approximating queries on sub-cubes from higher-level aggregations (e.g. [14]). However, these works did not have to cope with information that has been intentionally distorted.

Kargupta et al. used a data refurbishing approach to derive private information from a disguised data set [1]. Namely, a new data set *X*¤ is reconstructed from the disguised data using certain algorithms, and the difference between *X*¤ and the actual original data set *X* indicates how much private information can be disclosed. The further apart *X*¤ is from *X*, the higher level of the privacy preservation is achieved. Therefore, the difference between *X*¤ and *X* can be used as the measure to quantify how much privacy is preserved.

IV. TAXONOMY OF REFURBISHING-BASED PRIVACY PRESERVING TECHNIQUES

There are countless approaches which have been adopted for Privacy Preserving Data Mining . A variety of information can escort to the disclosure of private information in a disguised data set .We can classify them based on the following dimensions:

- Data correlations
- Data samples
- Partial or full value disclosure
- Data mining results
- Users or data miners
- Nature of data

A. Attribute or Feature Dependency

Attributes in many data sets are not self-governing, and some attributes might have a brawny correlationship amongst themselves. It is imperative to comprehend how such relationship can reason private information disclosure. The existing proposed two data refurbishing methods that are based on data correlations are - One method uses the Principal Component Analysis (PCA) technique, and the other method uses the Bayes Estimate (BE) technique.[25]

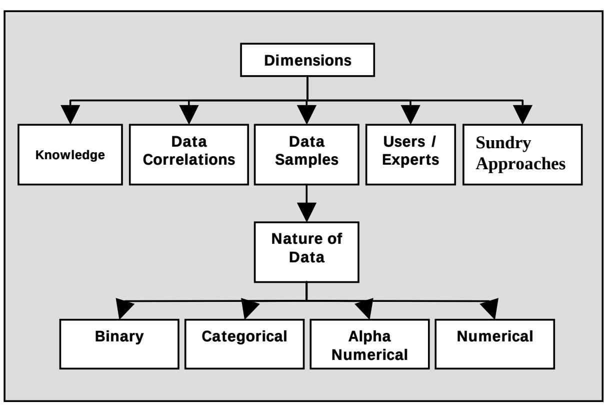


Figure 1. Taxonomy of Refurbishing based PPDM

B. Sample Dependency

For certain types of data sets, such as the time series data, there exists sequential enslavement amongst the samples. Even after tormenting the data with random noise, this enslavement can still be recovered. For instance, various techniques are available from the signal processing literature to de-clatter the tainted signals. One interesting research problem is: for different types of data, what kind of dependency relationships will facilitate the adversaries reconstruct the original data?

C. Partial Value Disclosure

In practice, it is probable that the values of some attributes can be disclosed (via other channels). For example, assume we have a medical database that is disguised by randomization schemes. Knowing that the patient Helen has cancer and uterus trouble, we might be capable to estimate the other information about her. How to reckon privacy under these states of affairs?

D. Data Mining Results

In the SMC approach, all the participating parties be able to see the concluding results. These results hold collective information about the data, which can escort to probable privacy breaches. For example, in the association rule mining, assume that there is a rule saying that *X* implies *Y* with 90% of support. Even if one party knows only *X* and the association rule results, he or she will be able to infer *Y* with high confidence. How do various data mining results, including classification models, association rules, and clustering impinge on individual privacy? Kantarcioglu et al. has initiated studies on this issue [1].

E. Users or data miners

- Users are not equally protective of all values in their records. Thus, users may be willing to provide modified values of certain fields by the use of a (publicly known) perturbing random distribution. This modified value may be generated using a custom code or a browser plug-in.

- Data mining problems do not necessarily require individual records, but only distributions. Since the perturbing distribution is known, it can be used to reconstruct *aggregate* distributions, i.e. the probability distribution of the data set. In many cases, data mining algorithms can be developed which use the probability distributions rather than individual records. An example of a classification algorithm which uses such aggregate information is discussed in [24].

F. Nature of data

A number of recently proposed techniques tackle the issue of privacy preservation by tormenting the data and reconstructing the distributions at an aggregate level in order to perform the mining. Beneath, we catalogue and classify some of these techniques based on the nature of data as.

1. Refurbishing-Based Techniques for Numerical Data

The effort presented in [36] addresses the quandary of building a decision tree classifier from exercise data in which the values of individual records have been perturbed. While it is impractical to perfectly guesstimate original values in individual data records, the authors propose a refurbishing modus operandi to precisely estimate the distribution of original data values. By using the reconstructed distributions, they are able to build classifiers whose accuracy is comparable to the accuracy of classifiers built with the original data. For the deformation of values, the authors have considered a discretization approach and a value deformation approach. For reconstructing the original distribution, they have considered a Bayesian approach and they proposed three algorithms for building accurate decision trees that rely on reconstructed distributions.

The work presented in [7] proposes an enhancement over the Bayesian-based refurbishing procedure by using an Expectation Maximization (EM) algorithm for distribution refurbishing. Further particularly, the authors confirm that the EM algorithm converges to the maximum likelihood estimate of the original distribution based on the perturbed data. They also illustrate that when a large amount of data is available, the EM algorithm provides stout estimates of the original distribution. It is also shown, that the privacy estimates of [34] had to be lowered when the additional knowledge that the miner obtains from the reconstructed aggregate distribution was included in the problem formulation.

2. Refurbishing-Based Techniques for Binary and Categorical Data

The exertion offered in [32] and [3] deal with binary and categorical data in the perspective of association rule mining. Both papers deem randomization techniques so as to offer privacy while they sustain high utility for the data set.

G. Sundry approaches

The work presented in [13] addresses the problem of building a decision tree classifier from training data in which the values of individual records have been perturbed. While it is not possible to accurately estimate original values in individual data records, the authors propose a refurbishing procedure to accurately estimate the distribution of original data values. By using the reconstructed distributions, they are able to build classifiers whose accuracy is comparable to the accuracy of classifiers built with the original data. For the distortion of values, the authors have considered a discretization approach and a value distortion approach. For refurbishing the original distribution, hey have considered a Bayesian approach and they proposed three algorithms for building accurate decision trees that rely on reconstructed distributions.

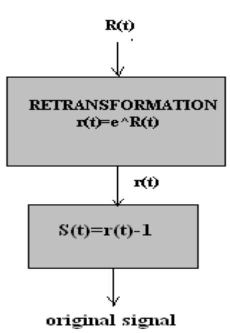
V. PRIVACY PRESERVING REFURBISHING-BASED TECHNIQUES

A. Notation for refurbishing

The goal of all mystification based approaches is that certain aggregate characteristics like mean and covariance matrices for numerical data, or marginal totals in contingency table data ought to remain basically unaffected though individual observations are befuddled. . Let b = x + m, where x and b represent the original and the befuddled data respectively while m represents type of mystification. All the existing mystification based approaches assume the estimated distribution F'_x is reconstructed from befuddled data should be the same or close to the original data distribution, F_x. Agarwal et al., pointed out that the (aggregate) reconstruction of the attribute values possibly will present a certain level of knowledge which can be used to deduce a data value to a superior level of accuracy and projected a privacy metric which takes into account this fact.

B. Process for refurbishing based PPDM

A reconstruction procedure reconstructs a new released dataset from the sanitized itemset lattice. Given the type of mystification added to the data, and the mystified data set, we must be able to reconstruct the original distribution (but not actual data values) of the data set. This facilitated a data mining algorithm to build a much more precise data than mining the mystified data alone.



VI. EVALUATION OF PRIVACY PRESERVING REFURBISHING-BASED TECHNIQUES

An important facet in the development and assessment of Refurbishing-Based Techniques and tools, for Privacy Preserving Data Mining is the identification of suitable evaluation criteria and the development of related benchmarks. It is often the case that no privacy preserving Refurbishing-Based Techniques exists that outperforms all the others on all possible criteria. Rather, Refurbishing-Based Techniques may perform better that another one on particular criteria, such as performance and/or data utility.

A preliminary list of evaluation parameters to be used for assessing the quality of Privacy Preserving Data Mining Refurbishing-Based Techniques is given below, based on different angles:

- Statistical Quality
- Quantification of Privacy and Information Loss
- Performance, Data Utility, Level of Uncertainity,Resistance

A. The statistical quality is measured in terms of bias, precision, and consistency [24].

1. **Bias** represents the difference between the unperturbed statistics and the expected value of its perturbed estimate.
2. **Precision** refers to the variance of the estimators obtained by the users.
3. **Consistency** represents the lack of contradictions and paradoxes. An exact disclosure occurs if by issuing one or more queries, a user is able to determine the exact value of a confidential attribute of an individual. A partial disclosure occurs if a user is able to obtain an estimator whose variance is below a given threshold.

B. Quantification of Privacy

The quantity used to measure privacy should indicate how closely the original value of an attribute can be estimated. The work in [16] uses a measure that defines privacy as follows:

> If the original value can be estimated with *c*% confidence to lie in the interval [.1; .2], then the interval width (.2 € .1) defines the amount of privacy at *c*% confidence level. For example, if the perturbing additive is uniformly distributed in an interval of width 2., then . is the amount of privacy at confidence level 50% and 2. is the amount of privacy at confidence level 100%.

Quantification of information loss

Given the perturbed values $z_1; z_2; : : : ; z_N$, it is (in general) not possible to reconstruct the original density function $fX(x)$ with an arbitrary precision. The greater the variance of the perturbation, the lower the precision in estimating $fX(x)$. We refer the lack of precision in estimating $fX(x)$ as information loss. In this section, we will consider how to quantify information loss. Note that the work in [1] uses an application dependent approach to measure the information loss. For example, for a classification problem, the inaccuracy in distribution refurbishing is measured by examining the effects on the misclassification rate.

C. Performance, data utility, level of uncertainty, resistance

A preliminary list of evaluation parameters to be used for assessing the quality of Privacy Preserving Data Mining Refurbishing-Based Techniques is given below:

Concerns: Performance, Scalability, Data Utility, Level of Uncertainty, Resistance

• the *performance* of the proposed Refurbishing-Based Techniques in terms of time requirements, that is the time needed by each Refurbishing-Based Techniques to hide a specified set of sensitive information;
• the *data utility* after the application of the privacy preserving technique, which is equivalent

with the minimization of the information loss or else the loss in the functionality of the data;

• the *level of uncertainty* with which the sensitive information that have been hidden can still be predicted;

• the *resistance* accomplished by the privacy Refurbishing-Based Techniques, to different data mining techniques.

It is thus imperative to endow the users with a set of metrics which will permit them to select the most appropriate privacy preserving technique for the data at hand; with respect to some specific parameters they are fascinated in optimizing.

VII. CONCLUSION

Weighing against conventional data alteration methods, data refurbishing is a novel hopeful, but not passably scrutinized method, which is aggravated by the fact, which it is yet to be explored. We have offered nomenclature and a all-inclusive explanation of various Privacy Preserving Data Mining algorithms. The slog presented in here, indicates the incessantly escalating implication of researchers in the environs of securing sensitive data and knowledge from malevolent users. The conclusions that we have reached from reviewing this environs, apparent that privacy issues can be effectively painstaking only within the restrictions of certain data mining algorithms. The lack of ability to take a broad view the results for classes of categories of data mining algorithms might be a vacillating terrorization for disclosing information. We predict the projected elucidation will bear up the new refurbishing-based line of investigation pathway and toil fine according to the appraisal metrics including trouncing effects, data utility, and time performance.

REFERENCES

[1] Sanil, A. Karr, X. Lin, and J. Reiter. Privacy preserving regression modelling via distributed computation. In *10th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, Seattle, WA, USA, August 2004.

[2] Alexandre Evfimievski, Ramakrishnan Srikant, Rakesh Agrawal, and Johannes Gehrke, *Privacy preserving mining of association rules*, In Proceedings of the 8th ACM SIGKDDD International Conference on Knowledge Discovery and Data Mining (2002).

[3] B. Gilburd, A. Schuster, and R. Wolf. A new privacy model and association-rule mining algorithm for large-scale distributed environments. In *10th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, Seattle, WA, USA, August 2004.

[4]B. Pinkas. Cryptographic techniques for privacy-preserving data mining. *SIGKDD Explorations*, 4(2), December 2002.

[5] C. Clifton, M. Kantarcioglu, J. Vaidya, X. Lin, and M. Y. Zhu. Tools for Privacy Preserving Data Mining . *SIGKDD Explorations*, 4(2), December 2002.

[6]C. Faloutsos, H.V. Jagadish, and N.D. Sidiropoulos. Recovering information from summary data. In Proc. of the 23rd Int'l Conference on Very Large Databases, pages 36{45, Athens, Greece, 1997.

[7] C.T. Yu and F.Y. Chin. A study on the protection of statistical databases. In *Proc. A CM SIGMOD Int. Conf. Management of Data,* pages 169-181, 1977.
]

[8]Chong K. Liew, Uinam J. Choi, and Chang J. Liew. A data distortion by probability distribution. *ACM TODS*, 10(3):395-411, 1985.

[9] D. Barbara and M. Sullivan. Quasi cubes: Exploiting approximations in multidimensional databases. SIGMOD Record, 26(3):12{17, 1997.

[10]D. Barbara, W. DuMouchel, C. Faloutsos, P. J. Haas, J. M. Hellerstein, Y. Ioannidis, H. V.Jagadish, T. Johnson, R.Ng, V. Poosala, and K. Sevcik. The New Jersey Data Reduction Report. Data Engrg. Bull., 20:3{45, Dec. 1997.

[11] D. Dobkin, A.K. Jones, and R.J. Lipton. Secure databases: Protection against user influence. *ACM TODS*, 4(1):97-106, March 1979.

[12] D. Meng, K. Sivakumar, and H. Kargupta. Privacy sensitive bayesian network parameter learning. In *The Fourth IEEE International Conference on Data Mining(ICDM)*, Brighton, UK, November 2004.

[13] D.E. Denning, P.J. Denning, and M.D. Schwartz. The tracker: A threat to statistical database security. *ACM TODS*, 4(1):76-96, March 1979.

[14]D.E. Denning. *Cryptography and Data Security*. Addison-Wesley, 1982.

[15]D.E. Denning. Secure statistical databases with random sample queries. *ACM TODS*, 5(3):291-315, Sept. 1980.

[16]Dakshi Agrawal and Charu C. Aggarwal, *On the design and quantification of Privacy Preserving Data Mining algorithms*, In Proceedings of the 20th ACM Symposium on Principles of Database Systems (2001), 247–255.

[17]Data Mining, National Security, Privacy and Civil Liberties, Bhavani Thuraisingham, The National Science Foundation.

[18] E. Lefons, A. Silvestri, and F. Tangorra. An analytic approach to statistical databases. In *9th Int. Conf. Very Large Data Bases*, pages 260-- 274. Morgan Kaufmann, Oct-Nov 1983.

[19] F.Y. 12hin and G. Ozsoyoglu. Auditing and infrence control in statistical databases. *IEEE Trans. Softw. Eng.*, SE°8(6):113-139, April 1982.

[20]H.W. Engl, M. Hanke, and A. Neubaue. Regularization of Inverse Problems. Kluwer, 1996.

[21]I.P. Fellegi. On the question of statistical confidentiality: : *J. Am. Star. Assoc.*, 67(337):7- 18, March 1972.

[22]J. Vaidya and C. Clifton. Privacy-preserving k-means clustering over vertically partitioned data. In *Proceedings of the 9th ACM SIGKDD International Conference on Knowledge Discovery and Ddata Mining*, 2003.

[23]J. Vaidya and C. Clifton. Privacy preserving association rule mining in vertically partitioned data. In *Proceedings of the 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, July 23-26 2002.

[24] J.F. Traub, Y. Yemini, and H. Woznaikowski. The statistical security of a statistical database. *ACM TODS*, 9(4):672-679, Dec. 1984.

[25] K. Wang, P. Yu, and S. Chakraborty. Bottom-up generalization: A data mining solution to privacy protection. In *The*

Fourth IEEE International Conference on Data Mining(ICDM), Brighton, UK, November 2004.

[26]L.H. Cox. Suppression methodology and statistical disclosure control. *J. Am. Stat.* Assoc., 75(370):377-395, April 1980.

[27] Leland L. Beck. A security mechanism for statistical databases. *A CM TODS,* 5(3):316--338, September 1980.

[28] M. Kantarcioglu, J. Jin, and C. Clifton. When do data mining results violate privacy? In *Proceedings of the 10th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD 2004)*, Seattle, WA, USA, August 2004.

[29]Nabil R. Adam and John C. Wortman. Security control methods for statistical databases. *ACM Computing Surveys,* 21(4):515-556, Dec. 1989.

[30]R. Agrawal and R. Srikant. Privacy Preserving Data Mining . In *Proceedings of the ACM SIGMOD*, pages 439–450, 2000.

[31] R. Conway and D. Strip. Selective partial access to a database. In *Proc. ACM Annual Con].*,pages 85-89, 1976.

[32]R. Wright and Z. Yang. Privacy-preserving Bayesian network structure computation on distributed heterogeneous data. In *10th ACM SIGKDD.*

[33]Rakesh Agrawal and Ramakrishnan Srikant, *Privacy-preserving data mining*, In Proceedings of the ACM SIGMOD Conference on Management of Data (2000), 439–450.

[34]S. Goldwasser. Multi-party computations: Past and present. In *Proceedings of the 16th Annual ACM Symposium on Principles of Distributed Computing*, Santa Barbara, CA USA, August 21-24 1997.

[35] S.L. Warner. Randomized response: A survey technique for eliminating evasive answer bias. J. *Am. Star. Assoc.,* 60(309):63-69, March 1965.

[36]Shariq J. Rizvi and Jayant R. Haritsa, *Maintaining data privacy in association rule mining*, In Proceedings of the 28th International Conference on Very Large Databases (2002).

[37]Steven P. Reiss. Practical data-swapping: The first steps. *ACM TODS,* 9(1):20-37, 1984.

[38] W. Du, Y. S. Han, and S. Chen. Privacy-preserving multivariate statistical analysis: Linear regression and classification. In *Proceedings of the 4th SIAM International Conference on Data Mining*, Lake Buena Vista, Florida, USA, April 2004.

[39]Y. Lindell and B. Pinkas. Privacy Preserving Data Mining . In *Advances in Cryptology - Crypto2000, Lecture Notes in Computer Science*, volume 1880, 2000.

[40] Zhengli Huang, Wenliang Du and Biao Chen. Deriving Private Information from Randomized Data.In *SIGMOD* 2005 June 14-16, Baltimore, Maryland, USA ,2005.