Multilevel Privacy Preserving by Linear and Non-Linear Data Distortion

Sangore Rohidas Balu Shrikant Lade

IEEE

Abstract

These days privacy preservation topic is based on one of the heated topics of data mining today. With the development of data mining technology, an increasing number of data can be mined out to reveal some potential information about user. While this will lead to a severe problem, which is users' privacy may be violated easily. The goal of privacy preserving is to mine the potential valuable knowledge without leakage of sensitive records, in other words, use non-sensitive data to infer sensitive data. There are many research and branches in this area. Most of them analyze and optimize the technologies and algorithms of privacy preserving data mining. Privacy Preserving Data Mining (PPDM) is used to extract relevant knowledge from large amount of data and at the same time protect the sensitive information from the data miners. The problem in privacy-sensitive domain is solved by the development of the Multi-Level (Multi-Party) Trust Privacy Preserving Data Mining (MLT-PPDM) where multiple differently perturbed copies of the same data is available to data miners at different trusted levels. In MLT-PPDM data owners generate perturbed data by various techniques like Parallel generation, Sequential generation and On-demand generation. MLT-PPDM is robust against the diversity attacks.

Index Terms

Data mining, Data Perturbation, Multiparty Privacy Preserving.

1. INTRODUCTION

Data mining is defined as the non-trivial extraction of implicit, previously unknown, and potentially useful information from large databases. Advances in hardware technology have increased the capability to store and record personal data about consumers and individuals. Data mining is under attack from privacy advocates because of a misunderstanding about what it actually is and a valid concern about how it's generally done. This has caused concerns that personal data may be used for a variety of intrusive or malicious purposes. Privacy preserving data mining help to achieve data mining goals without scarifying the privacy of the individuals and without learning underlying data values. Privacy-preserving data mining (PPDM) refers to the area of data mining that seeks to safeguard sensitive information from unsolicited or unsanctioned disclosure.

2. PROBLEM FORMULATION

Every day dozens of electronic trails through various activities such as using credit cards, swapping security cards, talking over phones and using emails. Ideally, the data should be collected with the consent of the data subjects. The collectors should provide some assurance that the individual privacy will be protected. However, the secondary use of collected data is also very common. Secondary use is any use for which data were not collected initially. Additionally, it is a common practice that organizations sell the collected data to other organizations, which use these data for their own purposes.

Nowadays, data mining is a widely accepted technique for huge range of organizations. Organizations are extremely dependent on data mining in their everyday activities. The paybacks are well acknowledged and can hardly be overestimated. During the whole process of data mining these data, which typically contain sensitive individual information such as medical and financial information, often get exposed to several parties including collectors, owners, users and miners. Disclosure of such sensitive information can cause a breach of individual privacy. For example, the detailed credit card record of an individual can expose the private life style with sufficient accuracy.

Private information can also be disclosed by linking multiple databases belonging to giant data warehouses and accessing web data. The definition of privacy is context dependant. In some scenarios individual data values are private, whereas in other scenarios certain association or classification rules are private. For example, we consider a scenario where a health service provider releases its patient data set to facilitate research and general analyzes. They may consider sensitive attribute values belonging to an individual as private. A privacy protection technique should prevent a disclosure of such a sensitive attribute value. However, in another scenario two or more organizations decide to collaborate by releasing their data sets to each other for mining. An intruder or malicious data miner can learn sensitive attribute values such as disease type (e.g. HIV positive), and income (e.g. AUD 82,000) of a certain individual, through re-identification of the record from an exposed data set. We note that the removal of the

Manuscript received October 5, 2015
Manuscript revised October 20, 2015

names and other identifiers may not guarantee the confidentiality of individual records, since a particular record can often be uniquely identified from the combination of other attributes. Therefore, it is not difficult for an intruder to be able to re-identify a record from a data set if he/she has enough supplementary knowledge about an individual. It is also not unlikely for an intruder to have sufficient supplementary knowledge, such as ethnic background, religion, marital status and number of children of the individual.

3. Related Work

Theoretically, Yao's general purpose secure circuit-evaluation protocol [7] solves any distributed two-party privacy-preserving data mining problem. As a practical matter, how-
ever, the circuits for even megabyte-sized databases would be intractably large. The alternative has been to find secure special-purpose protocols for specific data mining problems.
These protocols typically use Yao's protocol for evaluation
g very small circuits in intermediate stages, but use other, more efficient, methods when examining the actual data.

Most results in privacy-preserving data mining assume that the data is either horizontally partitioned (that is, each party to the protocol has some subset of the rows of an imaginary "global database"), or vertically partitioned (that is, each party has some subset of the columns of the "global database") [6].
Privacy-preserving clustering has been previously addressed by Oliviera and Za¨iane [5], Vaidya and Clifton [10],. Oliviera and Za¨iane's work [5] use s data transformation in conjunction with partition-based and hierarchical clustering algorithms, while the others use cryptographic techniques to give privacy-preserving versions of the k-means clustering algorithm. Vaidya and Clifton's result [10] addresses privacy-preserving k-means clustering for vertically partitioned data, Jha, Kruger, and McDaniel's [09] addresses horizontally partitioned data, and Bunn and Ostrovsky [7] address arbitrarily-partitioned data.
Tzung Pei et al presented Evolutionary privacy preserving in data mining [12]. Collection of data, dissemination and mining from large datasets introduced threats to the privacy of the data. Some sensitive or private information about the individuals and businesses or organizations had to be masked
before it is disclosed to users of data mining. An evolutionary privacy preserving data mining method was proposed to find about what transactions were to be hidden from a database. Based on the preference and sensitivity of the individual's data
in the database different weights were assigned to the attributes of the individuals. The concept of prelarge item sets was used to minimize the cost of rescanning the entire database and speed up the evaluation process of chromosomes. The proposed approach was used to make a good tradeoff between privacy preserving and running time of the data mining algorithms.
Han and Keong Ng presented Privacy Preserving Genetic Algorithms for Rule Discovery [3]. Entire data set was partitioned between two parties, and genetic algorithm was used to find the best set of rules without publishing their actual private data. Two parties jointly developed fitness function to evaluate the results using each party's private data but not compromising the privacy of the data by Secure Fitness Evaluation Protocol. To meet the privacy related challenges, results generated by genetic algorithm were not compromising privacy of those two parities having partitioned data. Creation of initial population and ranking the individuals for reproduction were done jointly by both parties.

4. Background

Additionally, 94% of the respondents consider acquisition of their personal information by a business they do Protection Methods Privacy can be protected through different methods such as Data Modification and Secure Multi-party Computation. Privacy preserving techniques can be classified based on the protection methods used by them.

Data Perturbation

It is the most important technique in MLT-PPDM. It is a category of data modification approaches that protect the sensitive data contained in a dataset by modifying a carefully selected portion of attribute-values pairs of its transactions. The employed modification makes the released values inaccurate, thus protecting the sensitive data, but it also achieving preservation of the statistical properties of the dataset. The perturbation method used should be such that statistics computed on the perturbed dataset do not differ significantly from the statistics that would be obtained on the original dataset. Data perturbation approaches fall into two
main categories namely probability distribution approach and the value distortion approach. The probability distribution approach replaces the data with another sample from the same (estimated) distribution or by the distribution itself . On the other hand, the value distortion approach perturbs the values of data elements or attributes directly by some additive or multiplicative noise before it is released to the data miner

Data Modification technique

Data Modification techniques modify a data set before releasing it to the users. Data is modified in such a way that the privacy is preserved in the released data set, whereas the data quality remains high enough to serve the purpose of the release.

A data modification technique could be developed to protect the privacy of individuals, sensitive underlying patterns, or both. This class of techniques includes noise addition, data swapping, aggregation, and suppression.

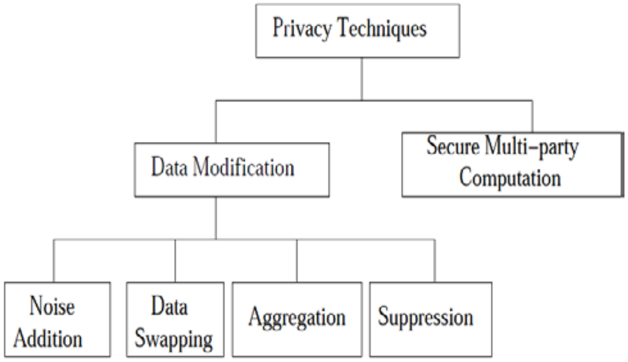

Figure 1. A Classification of Privacy Preserving Techniques.

Data Modification Existing privacy protection methods for centralized statistical databases can be categorized in three main groups, based on the approaches they take, such as query restriction, output perturbation, and data modification [1]. In a way, out of these privacy protection techniques, data modification is the most straightforward one to implement. Before the release of a data set for various data mining tasks and statistical analyses, it modifies the data set so that individual privacy can be protected while the quality of the released data remains high. After this modification of the data set we can use any off the shelf software such as DBMS, and See to manage and analyze the data without any restrictions on processing. That is not the case with query restriction and output perturbation. The simplicity of data modification techniques has made it attractive and widely used in the context of statistical database and data mining. Data modification can be done in many ways such as noise addition, data swapping, aggregation, and suppression.

Noise Addition in Statistical Database

Noise addition techniques were originally used for statistical databases which were supposed to maintain data quality in parallel to the privacy of individuals. Later on noise addition techniques were also found useful in privacy preserving data mining. The incorrectness in the statistic of a perturbed data set with respect to the statistic of the unperturbed data set is termed as bias. Mulalidhar et al. [14] presented a useful classification of various types of bias as follows.

- Bias due to the change in variance of an individual attribute.
- Bias due to the changes in relationship such as covariance, and correlation between confidential attributes.
- Bias due to the changes in relationship between confidential and nonconfidential attributes.
- Bias due to the change in the underlying distributions of a data set.

Different noise addition techniques, which we call Random Perturbation Technique (RPT), Probabilistic Perturbation Technique (PPT) and All Leaves Probabilistic Perturbation Technique (ALPT).

Data Swapping

Data swapping techniques were first devised by Dalenius and Reiss in 1982, for categorical values modification in the context of secure statistical databases [18]. The main appeal of the method was it keeps all original values in the data set, while at the same time makes the record re-identification very difficult. The method actually replaces the original data set by another one, where some original values belonging to a sensitive attribute are exchanged between them.

The noise is added to the class, i.e. the target attribute of a classifier, instead of all other attributes in the data set. As the class is typically a categorical attribute containing just two different values, the noise is added by changing the class in a small number of records. This is achieved by randomly shuffling the class attributes values belonging to heterogeneous leaves of a decision tree. If a leaf corresponds to a group of records having different class attribute values, then the leaf is known to be a heterogeneous leaf.

Aggregation

Aggregation is also known as generalization or global recoding. It is used for protecting an individual privacy in a released data set by perturbing the original data set prior to its release. Aggregation replaces k number of records of a data set by a representative record. The value of an attribute in such a representative record is generally derived by taking the average of all values, for the attribute, belonging to the records that are replaced. Due to the replacement of k number of original records by a representative record aggregation results in some information loss. The information loss can be minimized by clustering the original records into mutually exclusive groups of k records prior to aggregation. However, a lower information loss results in a higher disclosure risk since an intruder can make a better estimate of an original value from the attribute value of the released record. An adjustment of the cluster size i.e. the number of records in each cluster can produce an appropriate balance of information loss and disclosure risk [13].

Suppression

In suppression technique sensitive data values are deleted or suppressed prior to the release of a microdata. Suppression is used to protect an individual privacy from intruders' attempts to accurately predict a suppressed value. An intruder can take various approaches to predict a sensitive value. For example, a classifier, built on a released data set, can be used in an attempt to predict a suppressed attribute value. Therefore, sufficient number of attribute values should be suppressed in order to protect privacy.

Multi-Party Privacy Preserving Data Mining

The key goal in most distributed methods for privacy preserving data mining (PPDM) is to allow computation of useful aggregate statistics over the entire data set without compromising the privacy of the individual data sets within the different participants. Thus, the participants may wish to collaborate in obtaining aggregate results, but may not fully trust each other in terms of the distribution of their own data sets. For this purpose, the data sets may either be horizontally partitioned or be vertically partitioned. In horizontally partitioned data sets, the individual records are spread out across multiple entities, each of which has the same set of attributes. In vertical partitioning, the individual entities may have different attributes (or views) of the same set of records. Both kinds of partitioning pose different challenges to the problem of distributed privacy-preserving data mining.

5. Noise and Perturbation

Let G1 through GL be L Gaussian random variables. They are said to be jointly Gaussian if and only if each of them is a linear combination of multiple independent Gaussian random variables.[2]Equivalently, G1 through GL are jointly Gaussian if and only if any linear combination of them is also a Gaussian random variable.

A vector formed by jointly Gaussian random variables is called a jointly Gaussian vector. For a jointly Gaussian vector G = |G1, . . .;GL|T , its probability density function (PDF) is as follows: for any real vector g,

$$f_{\mathbb{G}}(g) = \frac{1}{\sqrt{(2\pi)^L \det(K_{\mathbb{G}})}} e^{-(g-\mu_{\mathbb{G}})^T K_{\mathbb{G}}^{-1} (g-\mu_{\mathbb{G}})/2},$$

where µG and KGG are the mean vector and covariance matrix of GG, respectively.

Corner-wave Property

Theorem 4 states that for M perturbed copies, the privacy goal in (10) is achieved if the noise covariance matrix KZZ has the corner-wave pattern as shown in (15). Specifically, we say that an M X M square matrix has the corner-wave property if, for every i from 1 to M, the following entries have the same value as the (i,i)th entry:

. all entries to the right of the (i, i) th entry in row i, and all entries below the (i, i) th entry in column i.

The distribution of the entries in such a matrix looks like corner-waves originated from the lower right corner.

Theorem 4. Let Y = |Y1T ; . . . ; YMT| represent an arbitrary number of perturbed copies. Assume that Y is generated from the original data X as follows:

Y = HX+Z

where H = [IN, . . . ; IN]^T , and Z = [Z$_1$^{T}, . . . ; Z$_M$^{T}] ^{T} with Zi $\sim N(0, \sigma_{Z_i}^2 K_X)$ is the noise vector. Without loss of generality, we

further assume

$$\sigma_{Z_i}^2 < \sigma_{Z_{i+1}}^2, \forall i = 1, \ldots, M-1$$

if Z is a jointly Gaussian vector and its covariance matrix KZ is given by

$$K_{\mathbb{Z}} = \begin{bmatrix} \sigma_{Z_1}^2 K_X & \sigma_{Z_1}^2 K_X & \cdots & \sigma_{Z_1}^2 K_X \\ \sigma_{Z_1}^2 K_X & \sigma_{Z_2}^2 K_X & \cdots & \sigma_{Z_2}^2 K_X \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{Z_1}^2 K_X & \sigma_{Z_2}^2 K_X & \cdots & \sigma_{Z_M}^2 K_X \end{bmatrix}.$$

This addition of noise is done for numeric dataset only as the dataset columns contain the numeric data that is added for generating the perturbed copy. But for all the different copy of the dataset let the value of σ is different then each perturbed copy is different from other. Although the perturbed copy are different from each other but if one get chance to collect few of them then there is the chance of producing the original data from the perturbed ones. perturbed copy of the data does not necessarily have more privacy since the added noise may be intelligently filtered out. Consequently, we define the privacy of a perturbed copy by taking into account an adversary's power in reconstructing the original data. We define the privacy of Y with respect to X to be D(X, X'(Y)), i.e., the distortion between X and the LLSE estimate X'(Y). A larger distortion hides the original values better, so we refer to a perturbed data Y2 to preserve more privacy than Y1 with respect to X if and only if D(X, X'(Y2) > D(X, X'(Y1)

Distortion

So for finding the perturbation between the datasets one term Distortion is introduce, To facilitate discussion on privacy, we define the concept of perturbation D between two data sets as the average expected square difference between them. For example, the distortion between the original data X and the perturbed copy Y = D +Z where D is the original copy and Z is the noise added.

$$\mathcal{D}(X,Y) = \frac{1}{N}\sum_{j=1}^{N} E[(y_j - x_j)^2] \geq 0.$$

It is easy to see that D(X, Y) = D(Y ,X). Based on the above definition, we refer to a perturbed copy Y2 to be more perturbed than Y1 with respect to X if and only if D(X, Y2) > D(X, Y1).

Linear Least Square Error

In order to generate original data from the perturbed copy linear least square method will reconstruct it. With the help of

$$X^{\wedge}(Y) = K_{XY}K_{Y}^{-1}(Y-\mu x)+ \mu x$$

Where Kx is the covariance of original data and Ky is covariance of perturbed data Y.

Multilevel Data Hierarchy

With the help of above calculation one can find the original data copy, so the purpose of perturbation is not fulfill. As if one get few perturbed copy of the original dataset then producing of the original is not a big task. So distributing perturbed copy of single level perturbation is not sufficient.

So instead of doing single level perturbation, multilevel perturbation is more fruit full as different level copy is distribute to the different user of different trust. This can be understand as the data owner decide the priority of the user for distributing the perturbed data copy. Now steps to improve the perturbation of the original copy is simple Let original copy is X which is perturbed to Y by adding noise Z.

$$Y = X + Z$$

Now for the lower level trust user new copy, is generate from the original data, then it will be not improve perturbation from the prior and the if higher level user can access the perturbed copy from the lower then chance of producing the original copy is more. So in order to reduce this probability of producing the original from the existing perturbed copy, perturbation for the new level is not generate from the original but it can be generate from the perturbed copy of the previous level.

So for first level

$$Y_1 = X + Z$$

For second level

$$Y_2 = Y_2 + Z'$$

For L level

$$Y_n = Y_{n-1} + Z''$$

Perturbation for Text Data

Here as the perturbation is done by adding noise generate by the Jointly Gaussian formula. So that would only perturbed the numeric data only. Because adding number to the number is possible but adding text to number is not possible so the perturbed copy would always contain the original data of each row in text form. Now in order to perturbed this copy for text data shuffling can be done. But this may produce the conflict for text and numeric mismatch. So in order to generate the noise in same sense, the proposed approach can perturbed the data in same sense as the numeric data is done. The proposed algorithm can perturbed both the numeric values as well as the text item set as well. So proper co-relation should be done between the text and numeric columns of the dataset.

6. Proposed Algorithm

In order to perturbed data for perturbation with full flexibility of generating any level of trusted copy at any time following argument one has to pass or decide for perturbation as privacy level L, original copy,

```
D ←Load()   // Dataset for perturbation
[N, A] ← ditinguise(D) //Separate into numeric and alphabetic data
Loop 1:L
    [M,C]←mean_covariance(N)
    Z←jointlty_gauss(N,M<C)
    //Following code is for text data perturbation
    Rand = max(Z)/type_of_item //
    Loop j = 1:count(N)
        Loop k = 1: type_of_item
            If Z(j)>Rand
                new_item = type_of_item(k)
            End-If
        End Loop
        new_row = D(i) + new_item  //here perturbation is done in ith transaction.
    End Loop
    Y = D + Z  // For numeric data direct addition
End Loop
```

In above algorithm

D is Dataset
N is numeric colum set
A is text colum set
L is level of the perturbed copy

M is mean of numeric colum
C is covariance of numeric colum
Z is noise produce by joit Gaussian method
Type_of_item is different item to be perturbed
New_item Item get select for perturbation
New_row is perturbed text row
Y is perturbed colum

7. Experiment and Results

In order to evaluate the above perturbation algorithm LLSE is used as the measuring parameter with the amount of distortion the is undertaken. An artificial dataset is produce for the experiment which include following columns {Transaction no, purchace_items, age, salary}. It is shown in below table.

Table 1: Representation of artificial dataset

Transaction_no	Purchace_items	Age	Salary
01	Jeans, T-Shirt	26	13000
02	T-shirt, Trouser	29	18040

So in the mention dataset the perturbation is done by adding noise in the numeric fields such as ‘age’ and ‘salary’ and the text fields are edit as per the algorithm and generated noise.

Table 2: Representation of artificial dataset after perturbation

Transaction_no	Purchace_items	Age	Salary
01	Jeans, T-Shirt, Trouser	27	13000
02	T-shirt, Trouser,	31	18040

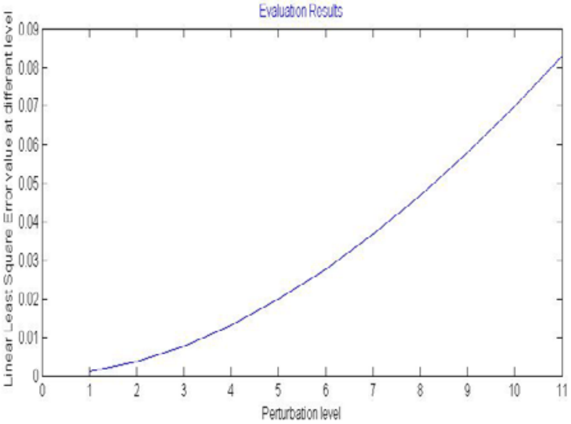

Figure 2: LLSE value at Perturbation level

LLSE Linear Least Square error LLSE value the smaller it is, the more accurate the LLSE estimation is. It generally decreases as more perturbed copies are used in the LLSE estimation.

LLSE = ((cov_x*(Y - mean(X))/cov_y)+ mean(X))
Where
cov_x is covariance of X , cov_y is covariance of Y

From above figure 2 of LLSE values at different perturbation level it can seen that by increasing the level LLSE vaue also increases which shows that, predicting of original values of numeric data is hard as the perturbation level increases.

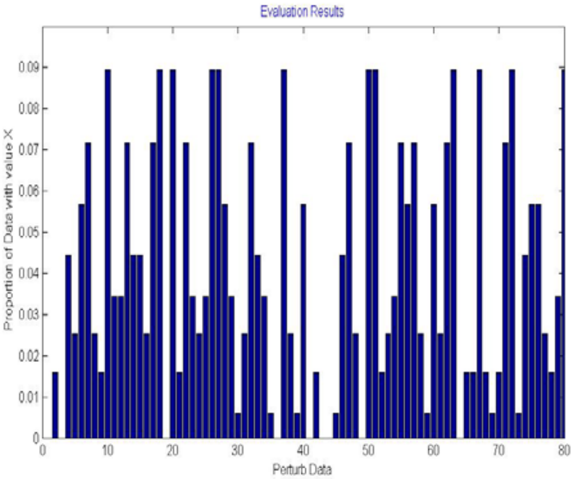

Figure 3:. Distribution of sensitive values Age

Above figure 3. represent the amount of perturbation done at any level in terms on the original data ratio. Bar value are random because of the joint noise values are different at different position. Although calculation of LLSE is only done for numeric values as the change in the text is not been predict by the LLSE formula or any kind of distortion function so method secure from any kind of data regeneration method.

Conclusion

Due to the right to privacy in the information ear, privacy-preserving data mining (PPDM) has become one of the newest trends in privacy and security and data mining research. In

this paper, this work introduced the related concepts of privacy-preserving data mining by developing multi party trust. Most of the work in this field is done either on text or on numeric but, this work use both numeric as well as text for perturbation of data. Results show that by decreasing the trust perturbation increases.

References

[1] J. K. Author, “Title of chapter in the book,” in Title of His Published Book, xth ed. City of Publisher, Country if not.

T zung -Pei, Hong Kuo-Tung Yang, Chun-Wei Lin and Shyue-Liang Wang, “Evolutionary privacy preserving in data mining “, IEEE World Automation Congress conference , 2010.

Shuguo Han Wee Keong Ng, “Privacy -Preserving Genetic Algorithms for Rule Discovery“, 2007.

Pei –Ling Chiu, “A Simulated Annealing Algorithm for General Threshold Visual Cryptography Schemes”, IEEE transactions, 2011.

S. Oliveira and O. R. Za¨iane. Privacy preserving clustering by data transformation. In Proc. Of the 18th Brazilian Symposium on Databases, pages 304–318, 2003.

Z. Yang and R. N. Wright. Privacy-preserving computation of bayesian networks on vertically partitioned data. IEEE Trans. on Knowledge and Data Engineering, 18(9):1253–1264, 2006.

A. C. Yao. How to generate and exchange secrets. In 27th FOCS, pages 162–167, 1986.

P. Bunn and R. Ostrovsky. Secure two-party k-means clustering. In
ACM Conference on Computer and Communications Security , pages 486–497, 2007

S. Jha, L. Kruger, and P. McDaniel. Privacy preserving clustering. In ESORICS , pages 397–417, 2005

J. Vaidya and C. Clifton. Privacy-preserving k-means clustering over vertically partitioned data. In 9th KDD, 2003.

T. Dalenius and S. P. Reiss. Data-swapping: A technique for disclosure control. Journal of Statistical Planning and Inference, 6(1):73{85, 1982.

V. Estivill-Castro and L. Brankovic. Data swapping: Balancing privacy against precision in mining for logic rules. In Proc. of Data Warehousing and Knowledge Discovery (DaWaK99), 1999.

Y. Li, S. Zhu, L.Wang, and S. Jajodia. A privacy-enhanced microaggregation method. In Proc. of 2nd International Symposium on Foundations of Information and Knowledge Systems, pages 148{159, 2002.

K. Muralidhar, R. Parsa, and R. Sarathy. A general additive data perturbation method for database security. Management Science, 45(10):1399{1415, 1999.

V. S. Verykios, E. Bertino, I. N. Fovino, L. P. Provenza, Y. Saygin, and Y. Theodoridis. State-of-the-art in privacy preserving data mining. SIGMOD Record, 33(1):50{57, 2004.