PRIVACY PRESERVING DATA MINING

A THESIS SUBMITTED IN PARTIAL FULFILLMENT
OF THE REQUIREMENTS FOR THE DEGREE OF

Bachelor of Technology

In

Computer Science and Engineering

By

Aman Jain

Roll No: 10306030

&

Bikash Sharma

Roll No: 10306005

Department of Computer Science and Engineering

National Institute of Technology

Rourkela

2007

PRIVACY PRESERVING DATA MINING

A THESIS SUBMITTED IN PARTIAL FULFILLMENT
OF THE REQUIREMENTS FOR THE DEGREE OF

Bachelor of Technology

In

Computer Science and Engineering

By

Aman Jain

Roll No: 10306030

&

Bikash Sharma

Roll No: 10306005

Under the guidance of

Prof. A.K. Turuk

Department of Computer Science and Engineering

National Institute of Technology

Rourkela

2007

National Institute of Technology
Rourkela

CERTIFICATE

This is to certify that the thesis entitled, "PRIVACY PRESERVING DATA MINING" submitted by Sri Bikash Sharma and Sri Aman Jain in partial fulfillment of the requirements for the award of Bachelor of Technology Degree in Computer Science and Engineering at the National Institute of Technology, Rourkela (Deemed University) is an authentic work carried out by them under my supervision and guidance.

To the best of my knowledge, the matter embodied in the thesis has not been submitted to any other university / institute for the award of any Degree or Diploma.

Date: 10 May, 2007

Prof A.K. Turuk
Dept. of Computer Science and Engineering
National Institute of Technology, Rourkela
Rourkela - 769008

ACKNOWLEDGEMENT

We express our sincere gratitude and indebtedness to **Dr. A.K. Turuk,** Professor in the department of Computer Science and Engineering, NIT Rourkela for giving us the opportunity to work under him and extending every support possible at each stage of this project work. The level of flexibility offered by him in implementing the project work is highly applaudable.

We would also like to convey our deep regards to all other faculty members and staff of Department of Computer Science and Engineering, NIT Rourkela, who have bestowed their whole hearted effort and guidance at appropriate times without which it would have been very difficult on our part to finish this project work.

Date: May 10, 2007 Aman Jain Bikash Sharma

CONTENTS

ABSTRACT

A fruitful direction for future data mining research will be the development of technique that incorporates privacy concerns. Specifically, we address the following question. Since the primary task in data mining is the development of models about aggregated data, can we develop accurate models without access to precise information in individual data records? We analyze the possibility of privacy in data mining techniques in two phases- randomization and reconstruction.

Data mining services require accurate input data for their results to be meaningful, but privacy concerns may influence users to provide spurious information. To preserve client privacy in the data mining process, techniques based on random perturbation of data records are used. Suppose there are many clients, each having some personal information, and one server, which is interested only in aggregate, statistically significant, properties of this information. The clients can protect privacy of their data by perturbing it with a randomization algorithm and then submitting the randomized version. This approach is called randomization. The randomization algorithm is chosen so that aggregate properties of the data can be recovered with sufficient precision, while individual entries are significantly distorted. For the concept of using value distortion to protect privacy to be useful, we need to be able to reconstruct the original data distribution so that data mining techniques can be effectively utilized to yield the required statistics.

Analysis

Let x_i be the original instance of data at client i. We introduce a random shift y_i using randomization technique explained below. The server runs the reconstruction algorithm (also explained below) on the perturbed value $z_i = x_i + y_i$ to get an approximate of the original data distribution suitable for data mining applications.

Randomization

We have used the following randomizing operator for data perturbation:

Given x, let R(x) be x+€ (mod 1001) where € is chosen uniformly at random in {-100…100}.

Reconstruction of discrete data set

P(X=x) = f X (x) ----Given

P(Y=y) = F y (y) ---Given

P (Z=z) = f Z (z) ---Given

f (X/Z) = P(X=x | Z=z)

= P(X=x, Z=z)/P (Z=z)

= P(X=x, X+Y=Z)/ f Z (z)

= P(X=x, Y=Z - X)/ f Z (z)

= P(X=x)*P(Y=Z-X)/ f Z (z)

= P(X=x)*P(Y=y)/ f Z (z)

Results

In this project we have done two aspects of privacy preserving data mining. The first phase involves perturbing the original data set using ‘randomization operator’ techniques and the second phase deals with reconstructing the randomized data set using the proposed algorithm to get an approximate of the original data set. The performance metrics like percentage deviation, accuracy and privacy breaches were calculated.

In this project we studied the technical feasibility of realizing privacy preserving data mining. The basic promise was that the sensitive values in a user’s record will be perturbed using a randomizing function and an approximate of the perturbed data set be recovered using reconstruction algorithm.

List of Figures

List of Tables

Chapter 1

INTRODUCTION

INTRODUCTION

Data mining and knowledge discovery in databases are two new research areas that investigate the automatic extraction of previously unknown patterns from large amounts of data. Data mining is concerned with the extraction of non-trivial, novel and potentially useful knowledge from large databases. Sequential data mining techniques have been applied successfully to a wide range of areas such as customer relationship management, web mining, science, engineering and medicine. However, a need for distributed and parallel data mining techniques has emerged over the past years. Data mining research deals with the extraction of potentially useful information from large collections of data with a variety of application areas such as customer relationship management, market basket analysis, and bioinformatics. The extracted information could be in the form of patterns, clusters or classification models. Association rules in a supermarket for example could describe the relationship among items bought together. Customers could be clustered in segments for better customer relationship management. Classification models could be built on customer profiles and shopping behavior to do targeted marketing. Many security and counter-terrorism-related decision support applications need data mining techniques for identifying emerging behavior, link analysis, building predictive models, and extracting social networks. They often deal with multi-party databases/data-streams where the data are privacy sensitive. Financial transactions, health-care records, and network communication traffic are a few examples. The power of data mining tools to extract hidden information from large collections of data lead to increased data collection efforts by companies and government agencies. Naturally this raised privacy concerns about collected data. In response to that, data mining researchers started to address privacy concerns by developing special data mining techniques under the framework of privacy preserving data mining. Opposed to regular data mining techniques, privacy preserving data mining can be applied to databases without violating the privacy of individuals. Recent advances in data collection, data dissemination and related technologies have inaugurated a new era of research where existing data mining algorithms should be reconsidered from a different point of view, this of privacy preservation. Privacy preserving data mining is a novel research direction in data mining and statistical databases, where data mining results are analyzed for the side-effects they

incur in data privacy. The main consideration in privacy preserving data mining is two fold. First, sensitive raw data should be modified or trimmed out from the original database, in order for the recipient of the data not to be able to compromise privacy. Second, sensitive knowledge which can be mined from a database by using data mining algorithms should also be excluded. The main objective in privacy preserving data mining is to develop algorithms for modifying the original data in some way, so that the private data and knowledge remain private even after the mining process. In a nutshell, the privacy preserving mining methods modify the original data in some way, so that the privacy of the user data is preserved and at the same time the mining models can be reconstructed from the modified data with reasonably accuracy. Various approaches have been proposed in the existing literature for privacy-preserving data mining which differ with respect to their assumptions of data collection model and user privacy requirements. The perturbation approach used in random perturbation model works under the strong privacy requirement that even the dataset forming server is not allowed learning or recovering precise records. There has been some research considering how much information can be inferred, calculated or revealed from the data made available through data mining process, and how to minimize the leakage of information. Overall privacy preserving data mining is an emerging technology that can prognosticate future trends and behaviors which could help to make proactive and knowledge driven decisions which thus helps in making the business model more targeted.

Chapter 2

State of the art in Privacy Preserving Data Mining

(i) Overview
(ii) Classification of Privacy Preserving Techniques
(iii) Review of Privacy Preserving Algorithms
(iv) Evaluation of Privacy Preserving Algorithms
(v) Performance of the proposed algorithms
(vi) Data Utility
(vii) Uncertainty Level
(viii)Endurance of Resistance to different Data Mining techniques
(ix) Distributed Data Mining
(x) Privacy Preserving Frequent Itemset Mining

Overview

Data mining and knowledge discovery in databases are two new research areas that investigate the automatic extraction of previously unknown patterns from large amounts of data. Recent advances in data collection, data dissemination and related technologies have inaugurated a new era of research where existing data mining algorithms should be reconsidered from a different point of view, this of privacy preservation. It is well documented that this new without limits explosion of new information through the Internet and other media, has reached to a point where threats against the privacy are very common on a daily basis and they deserve serious thinking. Privacy preserving data mining is a novel research direction in data mining and statistical databases, where data mining algorithms are analyzed for the side-effects they incur in data privacy. The main consideration in privacy preserving data mining is two fold. First, sensitive raw data like identifiers, names, addresses and the like should be modified or trimmed out from the original database, in order for the recipient of the data not to be able to compromise another person's privacy. Second, sensitive knowledge which can be mined from a database by using data mining algorithms should also be excluded, because such knowledge can equally well compromise data privacy, as we will indicate. The main objective in privacy preserving data mining is to develop algorithms for modifying the original data in some way, so that the private data and private knowledge remain private even after the mining process. The problem that arises when confidential information can be derived from released data by unauthorized users is also commonly called the "database inference" problem. In this report, we provide a classification and an extended description of the various techniques and methodologies that have been developed in the area of privacy preserving data mining.

The following figure gives the framework of Privacy preserving data mining.

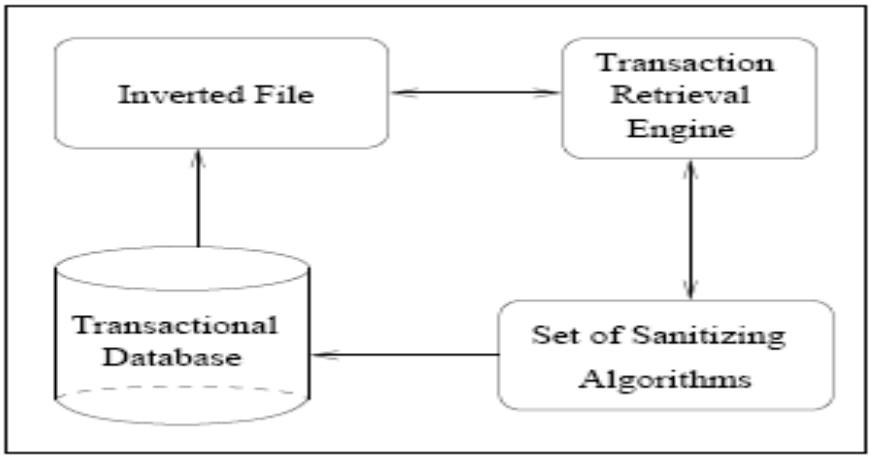

Fig: 2.1

2.1 Classification of Privacy Preserving Techniques

There are many approaches which have been adopted for privacy preserving data mining. We can classify them based on the following dimensions:

• Data distribution

• Data modification

• Data mining algorithm

• Data or rule hiding

• Privacy preservation

The first dimension refers to the distribution of data. Some of the approaches have been developed for centralized data, while others refer to a distributed data scenario. Distributed data scenarios can also be classified as horizontal data distribution and vertical data distribution. Horizontal distribution refers to these cases where different database records reside in different places, while vertical data distribution, refers to the cases where all the values for different attributes reside in different places. The second dimension refers to the data modification scheme. In general, data modification is used in order to modify the original values of a database that needs to be released to the public and in this way to ensure high privacy protection. It is important that a data modification

technique should be in concert with the privacy policy adopted by an organization. Methods of modification include:

• Perturbation, which is accomplished by the alteration of an attribute value by a new value (i.e., changing a 1-value to a 0-value, or adding noise),

• blocking, which is the replacement of an existing attribute value with a "?",

• Aggregation or merging which is the combination of several values into a coarser category,

• *S*wapping that refers to interchanging values of individual records, and

• Sampling, which refers to releasing data for only a sample of a population?

The third dimension refers to the data mining algorithm, for which the data modification is taking place. This is actually something that is not known beforehand, but it facilitates the analysis and design of the data hiding algorithm. We have included the problem of hiding data for a combination of data mining algorithms, into our future research agenda. For the time being, various data mining algorithms have been considered in isolation of each other. Among them, the most important ideas have been developed for classification data mining algorithms, like decision tree inducers, association rule mining algorithms, clustering algorithms, rough sets and Bayesian networks. The fourth dimension refers to whether raw data or aggregated data should be hidden. The complexity for hiding aggregated data in the form of rules is of course higher, and for this reason, mostly heuristics have been developed. The lessening of the amount of public information causes the data miner to produce weaker inference rules that will not allow the inference of confidential values. This process is also known as "rule confusion". The last dimension which is the most important refers to the privacy preservation technique used for the selective modification of the data. Selective modification is required in order to achieve higher utility for the modified data given that the privacy is not jeopardized.

The techniques that have been applied for this reason are:

• Heuristic-based techniques like adaptive modification that modifies only selected values that minimize the utility loss rather than all available values.

• Cryptography-based techniques like secure multiparty computation where a computation is secure if at the end of the computation, no party knows anything except its own input and the results, and

• Reconstruction-based techniques where the original distribution of the data is reconstructed from the randomized data.

It is important to realize that data modification results in degradation of the database performance. In order to quantify the degradation of the data, we mainly use two metrics. The first one, measures the confidential data protection, while the second measures the loss of functionality.

2.2 Review of Privacy Preserving Algorithms

2.2.1 Heuristic-Based Techniques

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

2.2.1.1 Centralized Data Perturbation-Based Association Rule Confusion

A formal proof that the optimal sanitization is an NP Hard problem for the hiding of sensitive large item sets in the context of association rules discovery. The specific problem which was addressed in this work is the following one. Let *D* be the source database, *R* be a set of significant association rules that can be mined from *D*, and let *Rh* be a set of rules in *R*. How can we transform database *D* into a database *D_*, the released database, so that all rules in *R* can still be mined from *D_*, except for the rules in *Rh*. The heuristic proposed for the modification of the data was based on data perturbation, and in particular the procedure was to change a selected set of 1-values to 0-values, so that the support of sensitive rules is lowered in such a way that the utility of the released database is kept to some maximum value. The utility in this work is measured as the number of non-sensitive rules that were hidden based on the side-effects of the data modification process. A subsequent work extends the sanitization of sensitive large item sets to the sanitization of sensitive rules. The approaches adopted in this work was either to prevent

the sensitive rules from being generated by hiding the frequent item sets from which they are derived, or to reduce the confidence of the sensitive rules by bringing it below a user-specified threshold. These two approaches led to the generation of three strategies for hiding sensitive rules. The important thing to mention regarding these three strategies was the possibility for both a 1-value in the binary database to turn into a 0-value and a 0-value to turn into a 1-value. This flexibility in data modification had the side-effect that apart from non-sensitive association rules that were becoming hidden; a non-frequent rule could become a frequent one. We refer to these rules as “ghost rules”. Given that sensitive rules are hidden, both non-sensitive rules which were hidden and non-frequent rules that became frequent (ghost rules) count towards the reduced utility of the released database. For this reason, the heuristics used for this later work, must be more sensitive to the utility issues, given that the security is not compromised.

2.2.1.2 Centralized Data Blocking-Based Association Rule Confusion

One of the data modification approaches which have been used for association rule confusion is data blocking. The approach of blocking is implemented by replacing certain attributes of some data items with a question mark. It is sometimes more desirable for specific applications (i.e., medical applications) to replace a real value by an unknown value instead of placing a false value. The introduction of this new special value in the data set imposes some changes on the definition of the support and confidence of an association rule. In this regard, the minimum support and minimum confidence will be altered into a minimum support interval and a minimum confidence interval correspondingly. As long as the support and/or the confidence of a sensitive rule lie below the middle in these two ranges of values, then we expect that the confidentiality of data is not violated. Notice that for an algorithm used for rule confusion in such a case, both 1-values and 0-values should be mapped to question marks in an interleaved fashion; otherwise, the origin of the question marks will be obvious.

2.2.1.3 Centralized Data Blocking-Based Classification Rule Confusion

In the classification rule framework, the data administrator has as a goal to block values for the class label. By doing this, the receiver of the information, will be unable to build informative models for the data that is not downgraded. Parsimonious downgrading is a framework for formalizing the phenomenon of trimming out information from a data set for downgrading information from a secure environment (it is referred to as High) to a public one (it is referred to as Low), given the existence of inference channels. In parsimonious downgrading a cost measure is assigned to the potential downgraded information that it is not sent to Low. The main goal to be accomplished in this work, is to find out whether the loss of functionality associated with not downgrading the data, is worth the extra confidentiality. Classification rules, and in particular decision trees are used in the parsimonious downgrading context in analyzing the potential inference channels in the data that needs to be downgraded. The technique used for downgrading is the creation of the so called parametric base set. In particular, a. parameter θ, $0 \leq \theta \leq 1$ is placed instead of the value that is blocked. The parameter represents a probability for one of the possible values that the attribute can get. The value of the initial entropy before the blocking and the value of the entropy after the blocking is calculated. The difference in the values of the entropy is compared to the decrease in the confidence of the rules generated from the decision tree in order to decide whether the increased security is worth the reduced utility of the data the Low will receive. The system is composed of a knowledge-based decision maker, to determine the rules that may be inferred, a “guard” to measure the amount of leaked information, and a parsimonious down grader to modify the initial downgrading decisions. The algorithm used to downgrade the data finds which rules from those induced from the decision tree induction, are needed to classify the private data. Any data that do not support the rules found in this way, are excluded from downgrading along with all the attributes that are not represented in the rules clauses. From the remaining data, the algorithm should decide which values to transform into missing values. This is done in order to optimize the rule confusion. The “guard” system determines the acceptable level of rule confusion.

2.2.1.4 Cryptography-Based Techniques

A number of cryptography-based approaches have been developed in the context of privacy preserving data mining algorithms, to solve problems of the following nature. Two or more parties want to conduct a computation based on their private inputs, but neither party is willing to disclose its own output to anybody else. The issue here is how to conduct such a computation while preserving the privacy of the inputs. This problem is referred to as the Secure Multiparty Computation (SMC) problem. In particular, an SMS problem deals with computing a probabilistic function on any input, in a distributed network where each participant holds one of the inputs, ensuring independence of the inputs, correctness of the computation, and that no more information is revealed to a participant in the computation than that's participant's input and output.

2.2.2 Reconstruction-Based Techniques

A number of recently proposed techniques address the issue of privacy preservation by perturbing the data and reconstructing the distributions at an aggregate level in order to perform the mining. Below, we list and classify some of these techniques.

2.2.2.1 Reconstruction-Based Techniques for Numerical Data

The work presented 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 reconstruction 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 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 proposes an improvement over the Bayesian-based reconstruction procedure by using an Expectation Maximization (EM) algorithm for distribution reconstruction. More specifically, the

authors prove that the EM algorithm converges to the maximum likelihood estimate of the original distribution based on the perturbed data. They also show that when a large amount of data is available the EM algorithm provides robust estimates of the original distribution. It is also shown, that the privacy estimates had to be lowered when the additional knowledge that the miner obtains from the reconstructed aggregate distribution was included in the problem formulation.

2.2.2.2 Reconstruction-Based Techniques for Binary and Categorical Data

The work presented deal with binary and categorical data in the context of association rule mining. Both papers consider randomization techniques that offer privacy while they maintain high utility for the data set.

2.3 Evaluation of Privacy Preserving Algorithms

An important aspect in the development and assessment of algorithms 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 algorithm exists that outperforms all the others on all possible criteria. Rather, an algorithm may perform better that another one on specific criteria, such as performance and/or data utility. It is thus important to provide users with a set of metrics which will enable them to select the most appropriate privacy preserving technique for the data at hand; with respect to some specific parameters they are interested in optimizing. A preliminary list of evaluation parameters to be used for assessing the quality of privacy preserving data mining algorithms is given below:

• *The performance* of the proposed algorithms in terms of time requirements, which is the time needed by each algorithm 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 has been hidden can still be predicted;

• The *resistance* accomplished by the privacy algorithms, to different data mining techniques.

Below we refer to each one of these evaluation parameters and we analyze them.

2.4 Performance of the proposed algorithms

A first approach in the assessment of the time requirements of a privacy preserving algorithm is to evaluate the computational cost. In this case, it is straightforward that an algorithm having an O (n^2) polynomial complexity is more efficient than another one with O (e^n) exponential complexity. An alternative approach would be to evaluate the time requirements in terms of the average number of operations, needed to reduce the frequency of appearance of specific sensitive information below a specified threshold. This values, perhaps, does not provide an absolute measure, but it can be considered in order to perform a fast comparison among different algorithms. The *communication cost* incurred during the exchange of information among a number of collaborating sites, should also be considered. It is imperative that this cost must be kept to a minimum for a distributed privacy preserving data mining algorithm.

2.5 Data Utility

The utility of the data, at the end of the privacy preserving process, is an important issue, because in order for sensitive information to be hidden, the database is essentially modified through the insertion of false information (swapping of values is a side effect in this case)or through the blocking of data values. We should notice here that some of privacy preserving techniques, like the use of sampling, do not modify the information stored in the database, but still, the utility of the data falls, since the information is not complete in this case. It is obvious that the more the changes are made to the database, the less the database reflects the domain of interest. Therefore, an evaluation parameter for the data utility should be the amount of information that is lost after the application of privacy preserving process. Of course, the measure used to evaluate the information loss depends on the specific data mining technique with respect to which a privacy algorithm is performed. For example, information loss in the context of association rule mining will be measured either in terms of the number of rules that were both remaining and lost

in the database after sanitization, or even in terms on the reduction/increase in the support and confidence of all the rules. For the case of classification, we can use metrics similar to those used for association rules. Finally, for clustering, the variance of the distances among the clustered items in the original database and the sanitized database can be the basis for evaluating information loss in this case.

2.6 Uncertainty Level

The privacy preservation strategies operate by downgrading the information that we want to protect below certain thresholds. The hidden information, however, can still be inferred even though with some uncertainty level. A sanitization algorithm then can be evaluated on the basis of the uncertainty that it introduces during the reconstruction of the hidden information. From an operational point of view, a scenario would be to set a maximum to the perturbation of information, and then consider the degree of uncertainty achieved by each sanitization algorithm under this constraint. We expect that the algorithm that will attain the maximum uncertainty level will be the one which will be preferred over all the rest.

2.7 Endurance of Resistance to different Data Mining techniques

The ultimate aim of hiding algorithms is the protection of sensitive information against unauthorized disclosure. In this case, it is important not to forget, that intruders and data terrorists will try to compromise information by using various data mining algorithms. Consequently, a sanitization algorithm developed against a particular data mining technique that assures privacy of information, may not attain similar protection against all possible data mining algorithms. In order to provide for a complete evaluation of sanitization algorithms, we need to measure its endurance against data mining techniques which are different from the technique that a sanitization algorithm has been developed for. We call such a parameter the *transversal endurance*. The evaluation of this parameter needs the consideration of a class of data mining algorithms which are significant for our test. Alternatively, we may need to develop a formal framework that upon testing of a sanitization algorithm against pre-selected data sets, we can transitively prove privacy assurance for the whole class of sanitization algorithms.

2.8 Distributed Data Mining

In contrast to the centralized model, the Distributed Data Mining (DDM) model assumes that the data sources are distributed across multiple sites. Algorithms developed within this field address the problem of efficiently getting the mining results from all the data across these distributed sources. Since the primary (if not only) focus is on efficiency, most of the algorithms developed to date do not take security consideration into account. With distributed data, the way the data is distributed also plays an important role in defining the problem. Data could be partitioned into many parts either vertically or horizontally.

2.8.1 Vertical Partitioning

Vertical partitioning (a.k.a. heterogeneous distribution) of data implies that though different sites gather information about the same set of entities, they collect different feature sets. For example, financial transaction information is collected by banks, while the IRS collects tax information for everyone.

The below figure gives a snapshot of vertically partitioned data.

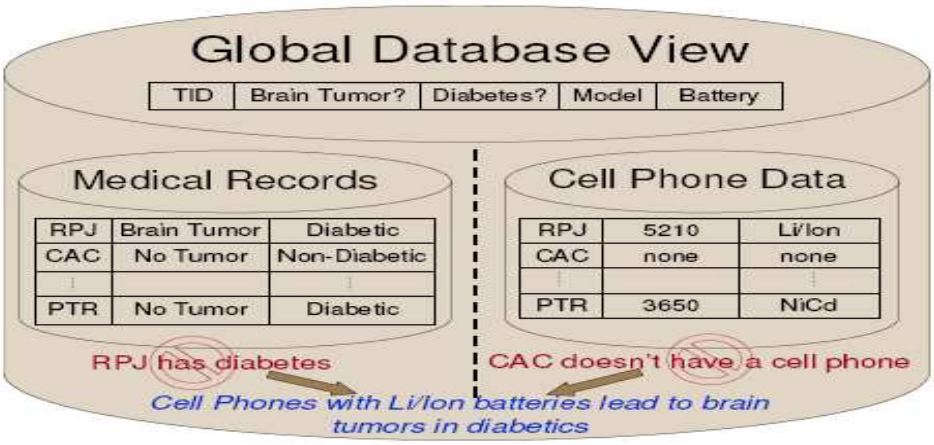

Fig 2.2: Vertical Partitioning

2.8.2 Horizontal Partitioning

In horizontal partitioning (a.k.a. homogeneous distribution), different sites collect the same set of information, but about different entities. An example of that would be grocery shopping data collected by different supermarkets (also known as market-basket data in the data mining literature). These different partitionings pose different problems, leading to different algorithms for privacy-preserving data mining.

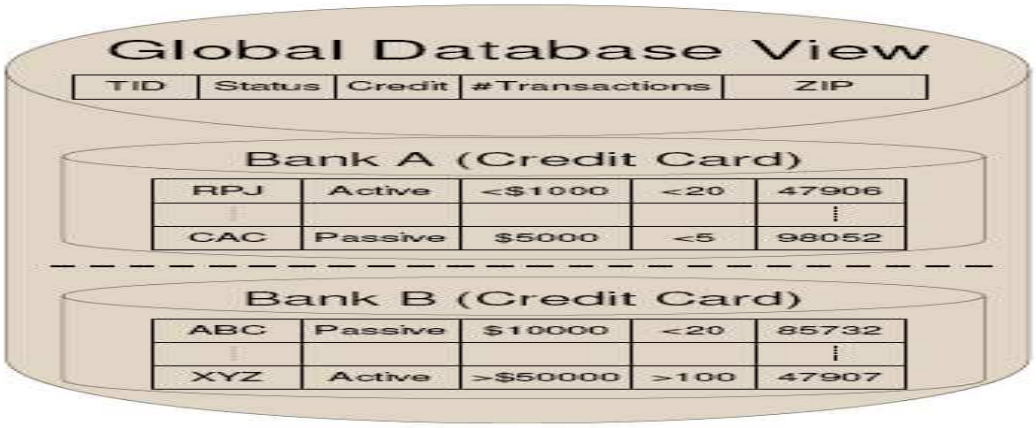

Fig 2.3: Horizontal Partitioning

2.9 Privacy Preserving Frequent Itemset Mining

Definition:

Let D be a transactional database, D be a distorted database from D in order to preserve individual privacy in D. It is this distorted database D* that is eventually supplied to the data miner, along with a description of the distortion procedure. The data miner mines the distorted database D* to estimate the frequent itemsets with support count satisfying the minimal support in the original database D, by virtue of the distribution procedure. Figure 2.4 illustrates the process.*

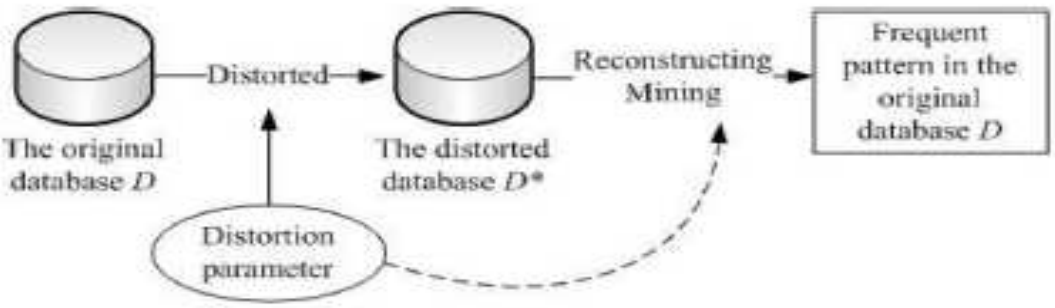

Fig 2.4: Privacy preserving frequent itemset mining process

Chapter 3

`

Randomization in Privacy Preserving Data Mining

(i) Background
(ii) Privacy preservation of continuous and discrete data
(iii) Randomization

3.1 Background

Randomization is an economical and efficient approach for privacy preserving data mining (PPDM). In order to guarantee the performance of data mining and the protection of individual privacy, optimal randomization schemes need to be employed. Suppose there are many clients, each having some personal information, and one server, which is interested only in aggregate, statistically significant, properties of this information. The clients can protect privacy of their data by perturbing it with a randomization algorithm and then submitting the randomized version. The randomization algorithm is chosen so that aggregate properties of the data can be recovered with sufficient precision, while individual entries are significantly distorted. How much distortion is needed to protect privacy can be determined using a privacy measure. Several possible privacy measures are known; finding the best measure is an open question. Methods and results in randomization for numerical and categorical data are much in focus. Suppose that some company needs to construct an aggregate model of its customer's personal data. For example, a retail store wants to know the age and income of its customers who are more likely to buy DVD players or mountain ski equipment; a movie recommendation system would like to learn users movie preferences in order to make advertisements more targeted; or an on-line business arranges its web pages according to an aggregate model of its website visitors. In all these cases, there is one central server (the company), and many clients (the customers), each having a piece of information. The server collects this information and builds its aggregate model using, for example, a classification algorithm for an algorithm for mining association rules. Often the resulting model no longer contains personally identifiable information, but contains only averages over large groups of clients. The usual solution to the above problem consists in having all clients send their personal information to the server. However, many people are becoming increasingly concerned about the privacy of their personal data. They would like to avoid giving out much more about themselves than is required to run their business with the company. If all the company needs is the aggregate model, a solution is preferred that reduces the disclosure of private data while still allowing the server to build the model. One possibility is as follows: before sending its piece of data, each client perturbs its own

share of information so that some true information is taken away and some false privacy information is introduced. This approach is called randomization. Another possibility is to decrease precision of the transmitted data by rounding, suppressing certain values, replacing values with intervals, or replacing categorical values by more general categories up the taxonomical hierarchy. The usage of randomization for preserving privacy has been studied extensively in the framework of statistical databases. In that case, the server has a complete and precise database with the information from its clients, and it has to make a version of this database public, for others to work with. One important example is census data: the government of a country collects private information about its inhabitants, and then has to turn this data into a tool for research and economic planning. However, it is not assumed that private records of any given person should not be released nor be recoverable from what is released. In particular, a company should not be able to match up records in the publicly released database with the corresponding records in the companies own database of its customers. In the case of statistical databases, however, the database is randomized when it is already fully known. This is different from our problem, where the randomization procedure is run on the client's side, and must be decided upon before the data is collected. A randomization for a statistical database is usually chosen so that it preserves certain aggregate characteristics (averages and covariance matrices for numerical data, or marginal totals in contingency tables for categorical data), or changes them in a predetermined way. Besides randomization, other privacy preserving transformations are used such as sampling and swapping values among records.

3.2 Privacy preservation of continuous and discrete data

Most security applications deal with heterogeneous data from different sources. This section considers some of the common data types that these applications usually deal with and discusses some of the existing random perturbation based privacy-preserving data mining algorithms for each of these domains. It first considers continuous valued data and a random data perturbation technique for privacy preservation of this type of

data. Next it considers discrete valued graph structured and transaction data for privacy-preserving applications.

3.2.1 Continuous Valued Data

Continuous valued data are widely prevalent among different data mining applications and security applications are no exceptions. Several randomized techniques have been proposed for privacy preserving data mining of continuous data. Random additive perturbation is one of them that is directly relevant to the work presented in this section. This section presents a brief review of this technique. It works by adding randomly generated noise from a given distribution to the values of sensitive attributes. The following sections discuss the data perturbation technique and the estimation of density functions from the perturbed data set.

3.2.2 Perturbing the data

The random additive perturbation method attempts to preserve privacy of the data by modifying values of the sensitive attributes using a randomized process.
There are two approaches: –
(1) Value Class Membership
(2) Value Distribution
In the more popular value distribution, the owner of a dataset returns a value u + v, where u is the original data, and v is a random value drawn from a certain distribution. Estimating the density function is a common problem in data mining and security applications are not an exception. The density information can be used for clustering, classification, and other related problems. Perturbed data using additive noise allows estimating the underlying density function reasonably well. Association rule learning is a widely popular technique for link analysis in data mining applications. Consider market basket transaction data in Boolean representations and a recently proposed randomized perturbation technique for privacy preserving association rule learning. Market basket data is usually a collection of transactions, where each transaction contains some product ids that are sold, and quantity sold. The transactions can be represented in a tabular form,

where each column represents one product id, and each row represents one transaction. For the sake of simplicity, let us consider transactions that only keep track of whether or not an item was purchased, not the quantity sold. In that case, we can represent a transaction using an l-dimensional Boolean string where l is the maximum number of different items that are available to a customer. The i-th bit will be set to 0 if the corresponding item is not sold in that transaction; it will be set to 1 otherwise. Therefore, one can represent a collection of m transactions using an m l dimensional Boolean matrix.

3.3 RANDOMIZATION

The problem of building classification models over randomized data was addressed. Each client has a numerical attribute, e.g. age, and the server wants to learn the distribution of these attributes in order to build a classification model. The clients randomize their attributes by adding random distortion values drawn independently from a known distribution such as a uniform distribution over a segment or a Gaussian distribution. The server collects the values and reconstructs the distribution using a version of the Expectation Maximization (EM) algorithm that provably converges to the maximum likelihood estimate of the desired original distribution. The goal is to discover association rules over randomized data. Each client has a set of items (called a *transaction*), e.g. product preferences, and here the server wants to determine all item sets whose support (frequency of being a subset of a transaction) is equal to or above a certain threshold. To preserve privacy, the transactions are randomized by discarding some items and inserting new items, and then are transmitted to the server. Statistical estimation of original supports and variances given randomized supports allows the server to adapt Apriori algorithm to mining item sets frequent in the non-randomized transactions by looking at only randomized ones.

Privacy:

It is not enough to simply concentrate on randomization and recovery of the model. We must also ensure that the randomization is sufficient for preserving *privacy*, as we randomized in the first place to achieve privacy. For example, suppose we randomize age x_i by adding a random number r_i drawn uniformly from a segment [-50, 50]. Assuming that the server receives age 120 from a user, privacy is somewhat compromised, as the server can conclude that the real age of the user cannot be less than 70 (otherwise $x_i + r_i <$ 70 + 50=120). Thus the server has learned a potentially valuable piece of information about the client information that is correct with 100% probability.

Example:

Suppose that private information x is a number between
0 and 1000. This number is chosen as a random variable X such that 0 is 1% likely whereas any non-zero is only about 0.1% likely

P[X=0] = 0.01

P[X=k] = 0.00099, k=1, 2, … ,1000

Suppose we want to randomize such a number by replacing it with
a new random number y=R(x) that retains some information about the original number x. Here are three possible ways to do it**:**

1. Given x, let $R_1(x)$ be x with 20% probability, and some other number (chosen uniformly at random) with 80% probability.
2. Given x, let $R_2(x)$ be x + € (mod 1001) where € is chosen uniformly at random in {-100,…, 100}.
3. Given x, let $R_3(x)$ be $R_2(x)$ be with 50% probability and a uniformly random number otherwise.

We can see that randomization operator R_1 reveals a lot of information about X when $R_1(x)$ happens to equal zero: the server learns with high probability that X originally was zero. Without knowing that $R_1(x) = 0$, the server considers X=0 to be just 1% likely; but

when $R_1(x) = 0$ is revealed, X=0 becomes about 70% likely. This does not happen when $R_2(x) = 0$ is revealed, the probability of X = 0 becomes only 4.8%.

Given:	$X = 0$	$X \notin \{200, \ldots, 800\}$
nothing	1%	$\approx 40.5\%$
$R_1(X) = 0$	$\approx 71.6\%$	$\approx 83.0\%$
$R_2(X) = 0$	$\approx 4.8\%$	100%
$R_3(X) = 0$	$\approx 2.9\%$	$\approx 70.8\%$

Table 1: Prior and posterior (given $R(X) = 0$) probabilities for properties in Example 1

Chapter 4

Reconstructing the Original data

(i) Reconstruction of the continuous data set

(ii) Reconstruction of the discrete data set

Concept

For the concept of using value distortion to protect privacy to be useful, we need to be able to reconstruct the original data distribution from the randomized data. Note that we reconstruct distributions, not values in individual records. We view the n original data values $x_1, x_2, \ldots\ldots, x_n$ of a one-dimensional distribution as realizations of n independent identically distributed random variables $X_1, X_2, \ldots\ldots, X_n$, each with the same distribution as the random variable X. To hide these data values, n independent random variables $Y_1, Y_2, \ldots\ldots, Y_n$ have been used, each with the same distribution as a different random variable Y. Given $x_1 + y_1$, $x_2 + y_2, \ldots\ldots, x_n + y_n$ and the cumulative distribution function F_Y of Y, we would like to estimate the cumulative distribution function F_X for X.

Reconstruction Problem:

Given a cumulative distribution function F_Y and the realizations of nd random samples $X_1 + Y_1, X_2 + Y_2, \ldots\ldots, X_N + Y_N$, estimate F_X.

Reconstruction of the continuous data set:

$$
\begin{aligned}
& F'_{X_1}(a) \\
& \quad \equiv \int_{-\infty}^{a} f_{X_1}(z \mid X_1 + Y_1 = w_1)\, dz \\
& \quad = \int_{-\infty}^{a} \frac{f_{X_1+Y_1}(w_1 \mid X_1 = z)\, f_{X_1}(z)}{f_{X_1+Y_1}(w_1)}\, dz \\
& \qquad \text{(using Bayes' rule for density functions)} \\
& \quad = \int_{-\infty}^{a} \frac{f_{X_1+Y_1}(w_1 \mid X_1 = z)\, f_{X_1}(z)}{\int_{-\infty}^{\infty} f_{X_1+Y_1}(w_1 \mid X_1 = z')\, f_{X_1}(z')\, dz'}\, dz \\
& \qquad \text{(expanding the denominator)} \\
& \quad = \frac{\int_{-\infty}^{a} f_{X_1+Y_1}(w_1 \mid X_1 = z)\, f_{X_1}(z)\, dz}{\int_{-\infty}^{\infty} f_{X_1+Y_1}(w_1 \mid X_1 = z)\, f_{X_1}(z)\, dz} \\
& \qquad \text{(inner integral is independent of outer)} \\
& \quad = \frac{\int_{-\infty}^{a} f_{Y_1}(w_1 - z)\, f_{X_1}(z)\, dz}{\int_{-\infty}^{\infty} f_{Y_1}(w_1 - z)\, f_{X_1}(z)\, dz} \\
& \qquad \text{(since } Y_1 \text{ is independent of } X_1\text{)} \\
& \quad = \frac{\int_{-\infty}^{a} f_{Y}(w_1 - z)\, f_{X}(z)\, dz}{\int_{-\infty}^{\infty} f_{Y}(w_1 - z)\, f_{X}(z)\, dz} \\
& \qquad \text{(since } f_{X_1} \equiv f_X \text{ and } f_{Y_1} \equiv f_Y\text{)}
\end{aligned}
$$

To estimate the posterior distribution function F'_X given $x_1 + y_1, x_2 + y_2, \ldots, x_n + y_n$, we average the distribution functions for each of the X_i.

$$F'_X(a) = \frac{1}{n} \sum_{i=1}^{n} F'_{X_i} = \frac{1}{n} \sum_{i=1}^{n} \frac{\int_{-\infty}^{a} f_Y(w_i - z)\, f_X(z)\, dz}{\int_{-\infty}^{\infty} f_Y(w_i - z)\, f_X(z)\, dz}$$

The corresponding posterior density function, f'_X is obtained by differentiating F'_X:

$$f'_X(a) = \frac{1}{n} \sum_{i=1}^{n} \frac{f_Y(w_i - a)\, f_X(a)}{\int_{-\infty}^{\infty} f_Y(w_i - z)\, f_X(z)\, dz} \qquad (1)$$

Given a sufficiently large number of samples, we expect f'_X in the above equation to be very close to the real density function f_X. However, although we know f_Y,[1] we do not know f_X. Hence we use the uniform distribution as the initial estimate f^0_X, and iteratively refine this estimate by applying Equation 1. This algorithm is sketched out in Figure 1.

Let $N(I_p)$ be the number of points that lie in interval I_p (i.e. number of elements in the set $\{w_i | w_i \in I_p\}$. Since $m(w_i)$ is the same for points that lie within the same interval,

$$f'_X(I_p) = \frac{1}{n} \sum_{s=1}^{k} N(I_s) \times \frac{f_Y(m(I_s) - m(I_p))\, f_X(I_p)}{\sum_{t=1}^{k} f_Y(m(I_s) - m(I_t))\, f_X(I_t)\, L_t}$$

Finally, let $\Pr'(X \in I_p)$ be the posterior probability of X belonging to interval I_p, i.e. $\Pr'(X \in I_p) = f'_X(I_p) \times L_p$. Multiplying both sides of the above equation by L_p, and using $\Pr(X \in I_p) = f_X(I_p) \times L_p$, we get:

$$\Pr'(X \in I_p) = \frac{1}{n} \sum_{s=1}^{k} N(I_s) \times \frac{f_Y(m(I_s) - m(I_p))\ \Pr(X \in I_p)}{\sum_{t=1}^{k} f_Y(m(I_s) - m(I_t))\ \Pr(X \in I_t)} \quad (3)$$

We can now substitute Equation 3 in step 3 of the algorithm (Figure 1), and compute step 3 in $O(m^2)$ time.[2]

Stopping Criterion With omniscience, we would stop when the reconstructed distribution was statistically the same as the original distribution (using, say, the χ^2 goodness-of-fit test [Cra46]). An alternative is to compare the observed randomized distribution with the result of randomizing the current estimate of the original distribution, and stop when these two distributions are statistically the same. The intuition is that if these two distributions are close to each other, we expect our estimate of the original distribution to also be close to the real distribution. Unfortunately, we found empirically that the difference between the two randomized distributions is not a reliable indicator of the difference between the original and reconstructed distributions.

Instead, we compare successive estimates of the original distribution, and stop when the difference between successive estimates becomes very small (1% of the threshold of the χ^2 test in our implementation).

Reconstruction of discrete data set

$P(X=x) = f_X(x)$ ----Given

$P(Y=y) = F_y(y)$ ---Given

$P(Z=z) = f_Z(z)$ ---Given

$f(X/Z) = P(X=x \mid Z=z)$

$= P(X=x, Z=z)/P(Z=z)$

$= P(X=x, X+Y=Z)/ f_Z(z)$

$= P(X=x, Y=Z - X)/ f_Z(z)$

$= P(X=x)*P(Y=Z-X)/ f_Z(z)$

$= P(X=x)*P(Y=y)/ f_Z(z)$

Chapter 5

RESULTS

The following were the results obtained after randomization and reconstruction for two uniform distributions – Poisson and Binomial:

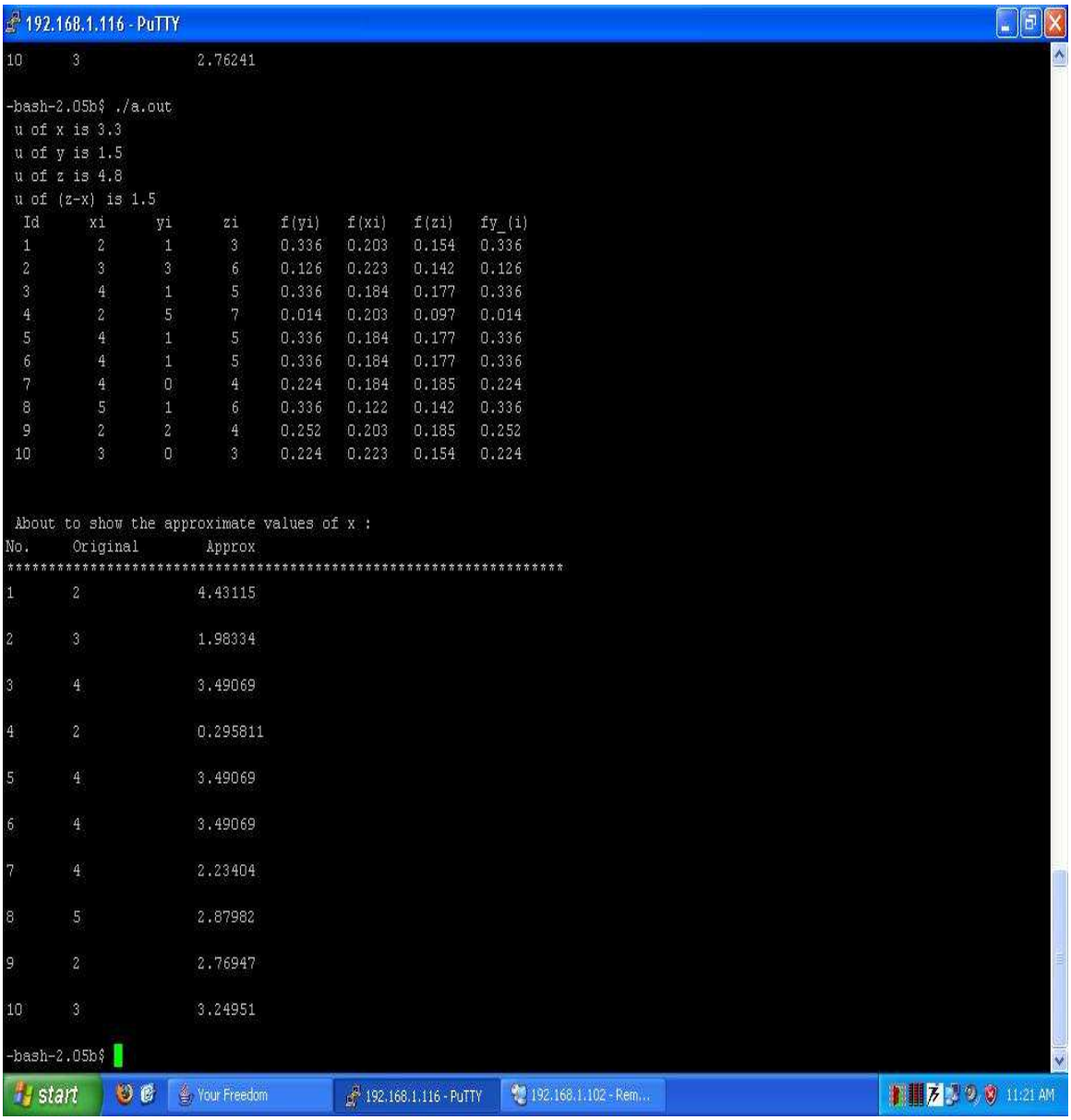

Fig 5.1: Poisson distribution

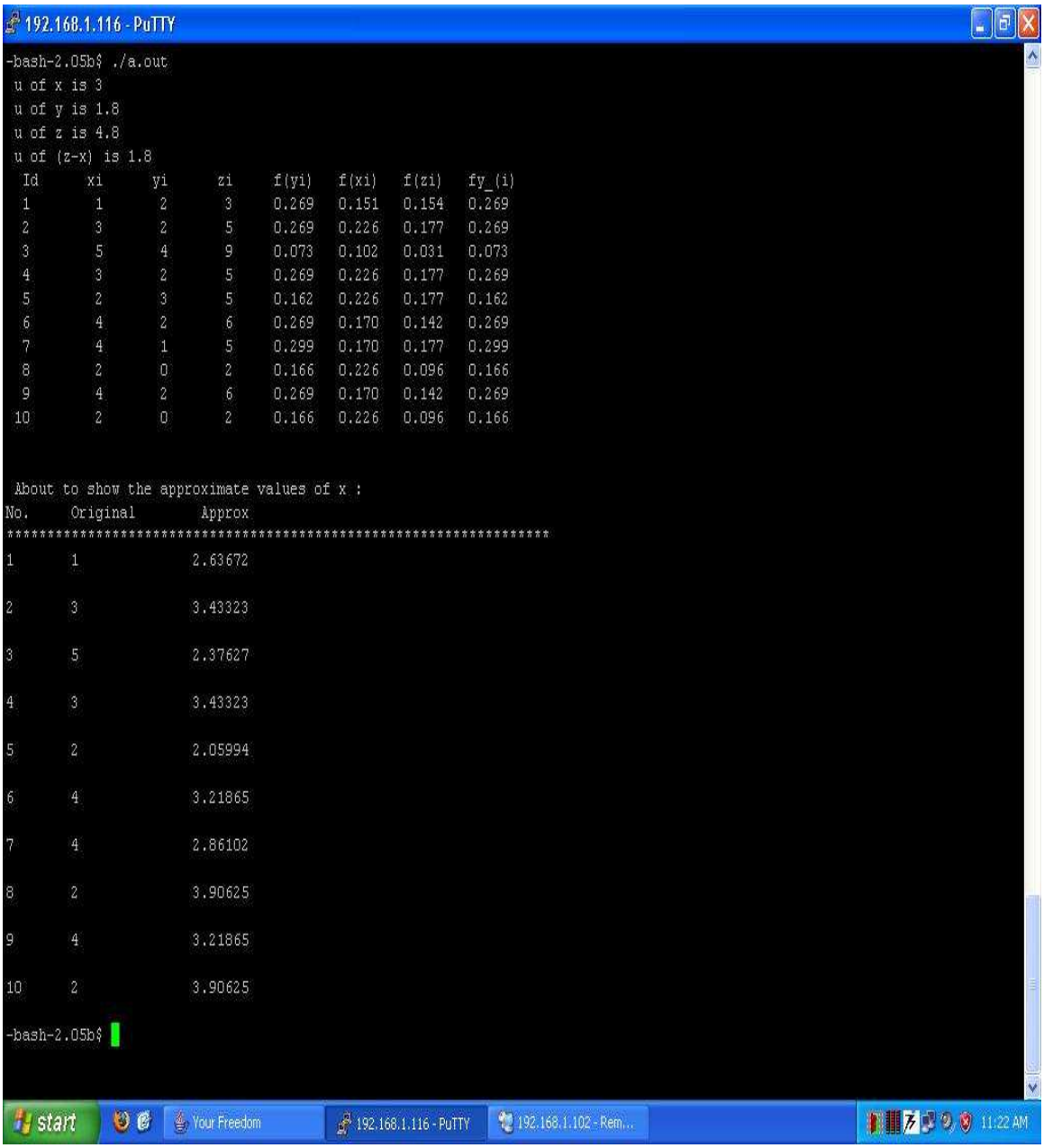

Fig 5.1: Binomial distribution

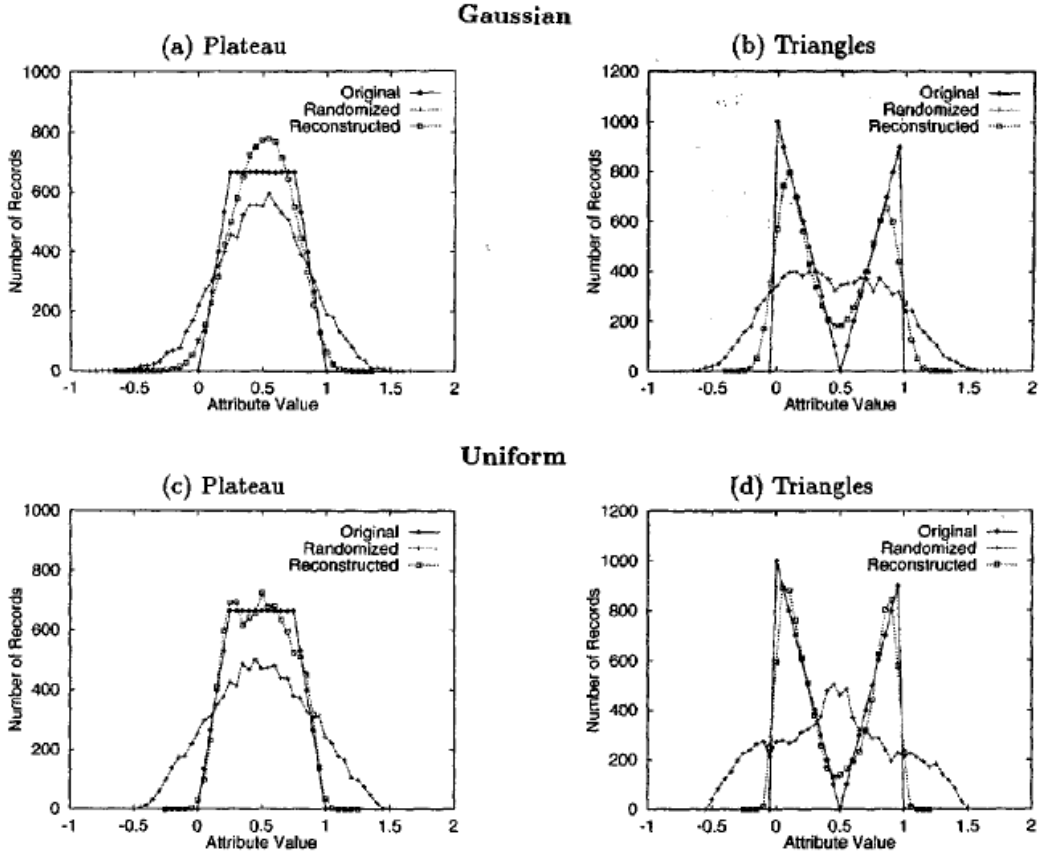

Fig 5.3: The variation of the original, randomized and reconstructed data for two distributions- Uniform and Gaussian.

Chapter 6

CONCLUSION

CONCLUSION

In this project, we studied the technical feasibility of realizing privacy preserving data mining. The basic premise was that the sensitive values in a user's record will be perturbed using a randomizing function so that they cannot be estimated with sufficient precision. Randomization can be done using Gaussian and Uniform perturbations. The question we addressed was whether given a large number of users who do this perturbation, can we still construct sufficiently accurate predictive models. We implemented data perturbation at the client's side using randomization operators and applied reconstruction algorithm at the server side to get an approximate of the client's original data set. The degree of distortion was assessed using some predefined performance metrics along with the extent of accuracy in the reconstruction.

REFERENCES

[1] R. Agrawal, T. Imielinski and A. Swami, "Mining association rules between sets of items in large databases", *Proc. of ACM SIGMOD Intl. Conference on Management of Data (SIGMOD)*, May 1993.

[2] R. Agrawal and R. Srikant, "Fast algorithms for mining association rules", *Proc. of 20th Intl. Conf. on Very Large Data Bases (VLDB)*, September 1994.

[3] R. Agrawal and R. Srikant, "Privacy-Preserving Data Mining", *Proc. of ACM SIGMOD Intl. Conf. on Management of Data*, May 2000.

[4] R. Agrawal, J. Kiernan, R. Srikant and Y. Xu, "Hippocratic Databases", *Proc. of 28th Intl. Conf. on Very Large Data Bases (VLDB)*, 2002.

[5] R. Agrawal, A. Kini, K. LeFevre, A. Wang, Y. Xu and D. Zhou, "Managing Healthcare Data Hippocratically", *Proc. of ACM SIGMOD Intl. Conf. on Management of Data*, 2004.

[6] R. Agrawal, R. Bayardo, C. Faloutsos, J. Kiernan, R. Rantzau and R. Srikant, "Auditing Compliance with a Hippocratic Database", *Proc. of 30th Intl. Conf. on Very Large Data Bases (VLDB)*, 2004.

[7] D. Agrawal and C. Aggarwal, "On the Design and Quantification of Privacy Preserving Data Mining Algorithms", *Proc. of Symposium on Principles of Database Systems(PODS)*, 2001.

[8] S. Agrawal, V. Krishnan and J. Haritsa, "On Addressing Efficiency Concerns in Privacy-Preserving Mining", *Proc. of 9th Intl. Conf. on Database Systems for Advanced Applications (DASFAA)*, March 2004.

[9] S. Agrawal and J. Haritsa, "A Framework for High-Accuracy Privacy-Preserving Mining", *Tech. Rep. TR-2004-02, DSL/SERC, Indian Institute of Science*, 2004. http://dsl.serc.iisc.ernet.in/pub/TR/TR-2004-02.pdf

[10] C. Aggarwal and P. Yu, "A Condensation Approach to Privacy Preserving Data Mining", *Proc. of 9th Intl. Conf. on Extending DataBase Technology (EDBT)*, March 2004

[11] M. Atallah, E. Bertino, A. Elmagarmid, M. Ibrahim and V. Verykios, “Disclosure Limitation of Sensitive Rules”, *Proc. of IEEE Knowledge and Data Engineering ExchangeWorkshop (KDEX)*, November 1999.

[12] Y. Censor, “Pareto Optimality in Multiobjective Problems”, *Appl. Math. Optimiz.*, vol. 4,1977.

[13] L.F. Cranor, J. Reagle, and M.S. Ackerman. “Beyond concern: Understanding net users’ attitudes about online privacy”. *Technical Report TR 99.4.3, AT&T Labs-Research*, April1999.

[14] Lorrie Faith Cranor, editor. *Special Issue on Internet Privacy*. Comm. A12M, 42(2), Feb.1999.

[15] Da Cunha, N.O. and E. Polak, “Constrained Minimization Under Vector-valued Criteria in Finite Dimensional Spaces,”*J. Math. Anal. Appl., Vol. 19, pp. 103-124,* 1967.

[16] E. Dasseni, V. Verykios, A. Elmagarmid and E. Bertino, “Hiding Association Rules by Using Confidence and Support”, *Proc. of 4th Intl. Information Hiding Workshop (IHW)*, April 2001.

[17] The Economist, “The End of Privacy”, May 1999.

[18] The European Union’s Directive on Privacy Protection, October 1998. Available from http://ww.echo.lu/legal/en/dataprot/ direct iv/direct iv. htlal.

[19] A. Evfimievski, R. Srikant, R. Agrawal and J. Gehrke, “Privacy Preserving Mining of Association Rules”, *Proc. of 8th ACM SIGKDD Intl. Conf. on Knowledge Discovery and Data Mining (KDD)*, July 2002.

[20] A. Evfimievski, J. Gehrke and R. Srikant, “Limiting Privacy Breaches in Privacy Preserving
Data Mining”, *Proc. of ACM Symp. on Principles of Database Systems (PODS)*,
June 2003.

[21] W. Feller, “An Introduction to Probability Theory and its Applications (Vol. I)”, Wiley,1988.