A
Dissertation
On

A Novel Approach for Privacy Preserving Data Mining Using Randomization

Submitted in Partial Fulfillment of the Requirement
For the Award of the Degree of

Master of Technology

In

Software Engineering

By

Ingle Pradeepkumar
University Roll No. 2K14/SWE/07

Under the Esteemed Guidance of

Dr. Rajni Jindal
Associate Professor
Computer Science and Engineering Department
Delhi Technological University

to
COMPUTER SCIENCE AND ENGINEERING DEPARTMENT

DELHI TECHNOLOGICAL UNIVERSITY

DELHI – 110042, INDIA

2016

DECLARATION

Date:…………..

I, **Ingle Pradeepkumar, University Roll No: 2K14/SWE/07**, hereby declare that the Major Project-II report titled "**A Novel Approach for Privacy Preserving Data Mining Using Randomization"** which is being submitted for the award of the degree of **Master of Technology** (**Software Engineering**) is a record of bonafide work carried out by me under the supervision of **Dr. Rajni Jindal, Associate Professor,** Department of Computer Science and Engineering, Delhi Technological University.

I further declare that the work presented in this report has not been submitted to any university or institution for the award of any diploma or degree.

Ingle Pradeepkumar
University Roll No: 2K14/SWE/07
M.Tech (Software Engineering)
Department of Computer Science and Engineering
Delhi Technological University

DELHI TECHNOLOGICAL UNIVERSITY

(Government of National Capital Territory of Delhi)

Main Bawana Road,

Delhi- 110042

CERTIFICATE

Date:……………….

This is to certify that the work embodied in the report titled **"A Novel Approach for Privacy Preserving Data Mining Using Randomization"** submitted to Delhi Technological University for the award of the degree of Master of Technology (Software Engineering) by **Ingle Pradeepkumar, University Roll No: 2K14/SWE/07** in the department of Computer Science and Engineering, Delhi Technological University is an authenticate work carried out by him under my guidance.

Project Superviser:

Dr. Rajni Jindal
Associate Professor
Department of Computer Science and Engineering
Delhi Technological University,
Delhi - 110042.

ACKNOWLEDGEMENT

First of all, I would like to express my deep sense of respect and gratitude to my project supervisor Dr. Rajni Jindal for providing the opportunity of carrying out this project and being the guiding force behind this work. I am deeply thankful to her for the support, advice and encouragement she provided without which the project could not have been a success.

Secondly, I am grateful to Dr. O.P.Verma, HOD, Computer Science and Engineering Department, DTU for his immense support. I would also like to acknowledge Delhi Technological University library and staff for providing the right academic resources and environment for this work to be carried out.

Last but not the least I would like to express sincere gratitude to my parents and friends for constantly encouraging me during the completion of work.

Ingle Pradeepkumar

University Roll no: 2K14/SWE/07

M.Tech (Software Engineering)

Department of Computer Science and Engineering

Delhi Technological University

ABSTRACT

Data mining is a knowledge extraction process from a huge amount of data. This data has many kinds such as pictorial data, analytical data, and survey data. There is information of individuals and organizations in the data to be mined. Many organizations have private and sensitive information about the individual. This sensitive information should not be publicized so that it can cause threat to the privacy of the individual. But for data mining purpose the information is needed to be publicized for various data mining tasks.

Privacy preserving data mining is a recent research subject for discovering new methods so that both the purpose of data mining should be fulfilled. Privacy of the individual is protected. Scope of this project is to study the recent research work on the privacy preserving data mining and improve or propose a new naval approach for privacy preserving data mining.

Among the techniques that are studied k-anonymization and randomization are used to get a new improved approach for privacy preserving data mining. In randomization method the matrix method is used to perturb the data. There are modifications that are made to the existing method. In k-anonymity generalization method is used. The combined method is proposed and analyzed so that it is efficient with respect to the concerned parameters.

CONTENT

List of Abbreviations

PPDM	Privacy Preserving Data Mining
SMC	Secure Multiparty Computation
ID	Identity
OLAP	Online Analytical Processing
RAM	Random Access Memory
IDE	Integrated Development Environment

List of Figures

List of Tables

CHAPTER 1

INTRODUCTION

1.1 Background

Data mining is now a familiar term as it has numerous applications. Data mining is basically used to get useful information from large amount of raw data. Many organizations collect individual information for their record. This information is used to do survey, analysis and statistic for knowledge discovery. The famous example is market basket analysis to understand the consumer behavior. Data mining has many applications in the 'Internet of Things'. Social networking sites have uses like sentiment analysis, browsing history analysis. Entertainment sites suggest the recommendations for the viewer based on its favorite which are available through the browsing history of that user. This data sometimes contains much personal and private information which reveals the identity of the person causing threat to privacy. In United States only 90% of the individual can be distinctly identified using the data they had given to the organizations (Sweeny, 2002). The personal information can be misused or it can be used for other purpose without permission. Thus the privacy of the person is intruded.

To solve above problem there are privacy preserving data mining techniques which intend to transform the data so that the identity of the individual should not be revealed at the same time the purpose of the data mining is accomplished (Du & Atallah, 2001). Privacy preserving data mining (PPDM) helps to support the data mining related operations. PPDM is very similar to Secure Multiparty Computation (SMC), which handles computing distributive. It is also a good research topic in the field of privacy preservation. PPDM is introduced to protect the private and sensitive information of an individual (Saranya, Premlatha & Rajasekar, 2015). Individual person would not like its personal information to be shared without his knowledge. For example, nobody wants his call details, medical records to be disclosed to an unauthorized way of access.

Many organizations share the data of the individuals among them for analysis and other purposes. The organization which collects such information should be aware of this and it is. That is why there is need for privacy preservation in the process of data mining. Mainly data mining is done for survey and statistical purpose. The privacy is nothing but the ability or right of the individual to control the private information about itself to full extent (Li, Yan & Zhang, 2014). The data should not be transformed much as it becomes unable to get the required knowledge. It should have its utility purpose. Thus data utility means how useful is the data for data mining. Data utility should not get affected.

PPDM concerns about how the privacy of the individual remains unaffected even after the various operations of data mining. PPDM techniques can be categorized on many grounds (Dhanalakshmi & Sankari, 2014). In one of the transformation they are categorized as data hiding and knowledge hiding techniques. The data hiding techniques have disadvantage of information loss. Data transformation techniques like perturbation, cryptography used so that the data can be retrieved from transformed data again. But these techniques are too simple to understand and fall prey to attacks. The perturbation technique is based on the method to add the noise to the original data. These techniques suffers through high overhead of noise addition also are simple to break. Cryptographic techniques can retrieve data back but it requires much computing power in terms of time and space. K-anonymity is the widely used method for data perturbation.

The existing procedures like k-anonymity, l-diversity, t-closeness are task dependent that is different algorithms are developed for different data mining tasks (Hellani, Kilany, & Sokhn, 2015). For example different algorithms may be required for association rule mining, clustering, and classification techniques. Randomization is also the most common method but it is also vulnerable to attacks. So combined methods are used to get more complexity. Randomization along with k-anonymity proves to be a better approach for achieving privacy in data mining.

1.2 Privacy Preservation in Data Mining

Privacy in data mining means how privacy is related to data mining. In hospitals patients' data is collected. These records have attributes like name, gender, age, disease, diagnose, ZIP code (Usha, Shriram, & Satishkumar, 2014). The election department also collects data about the citizens which include name, gender, birth date, address. In this data there are two types of identifiers:

- **Explicit Identifiers**: These identifiers can identify the individual independently. For example, name, ID number, social security number, Unique ID given to the citizen, etc.
- **Quasi Identifiers:** These identifiers can identify the individual if used combine. For example, age, gender, ZIP code, address, birth date, etc.

Privacy preserving data mining growing importance in recent decade as the people are concerned about it. There are different perspectives to how privacy is defined by researchers in the history. The main concern is to save the data from misuse. The data is of different types. Some firms collects records, some organizations have records of pictures. The data is collected for various reasons. To give public services to the citizens of the country, government collects data for identification and other information. The collected data is stored as records in the firm or it is out sourced to another firm.

For commercial, social, economic and various other reasons the data are analyzed. Useful information is extracted from it. The data to be processed is so huge, so powerful machines are used data mining tasks. The data which is collected at the organizations like banks, hospitals, universities have private, confidential, and sensitive information. This information, if disclosed causes threat to the privacy of the individual. So to protect the privacy of the individual there are many techniques which are proposed as to secure sensitive information along with fulfilling the data mining purpose with as efficiently as possible (Sharma et al.., 2014).

So here there are three things which we have to consider:

1. Privacy of the person
2. Data Mining purpose
3. Efficiency

The data should be transformed to such a level that the privacy of the person is protected at the same time the data mining purpose should be fulfilled. The literature review is done on such privacy preserving data mining techniques which are being researched recently. These techniques are evaluated on above parameters.

1.3 Randomization

Here, we will discuss the randomization method used for privacy preserving data mining. Randomization is a common and mostly used technique for privacy preserving data mining. In this technique the data is distorted using various methods. The method of distortion is called as perturbation (Zhang &Bi, 2010). Perturbation is process of adding noise to the existing original data. The perturbation is of two kinds:

1. **Additive Perturbation**
2. **Multiplicative Perturbation**

In additive perturbation random number is added to the original data or the data is incremented by random number. The table 2.1 and 2.2 shows the example of additive perturbation.

Table 1.1: Original data

Name	**Age**	**Account No**
Ram	43	78979863421
Shyam	34	78932487829
Jim	56	87239482983
Mary	23	78458978934

Table 1.2: Perturbed data

Name	**Age**	**Account No**
Ram	48(43+5)	78979863532(421+111)
Shyam	39	78932487940
Jim	61	87239483094
Mary	28	78458979045

Similarly, in multiplicative perturbation the original data is multiplied by a random constant. The level on which it is performed differ from the type of data set it is applied. The two methods are most suitable for numeric data.

Now we will describe the method of randomization as follows. Consider a set of data records denoted by $X = \{x_1...x_n\}$. Then for each element in X we add noise component from the probability distribution f_y of Y. The noise components are drawn independently and are denoted by $\{y_1 \ldots y_n\}$. Hence, the new set of distorted records becomes $x_1 + y_1 \; . \; . \; . \;\; x_n + y_n$. We describe the new set of records $z_1 . . . z_n$. The variance of the added noise is assumed to be large enough so that the original data could not be guessed from the distorted data. Thus, it is known that the original records cannot be recovered at the same time the distortion can be recovered from the data records.

Thus, if X denotes the element from the original set of data records and Y denotes the noise element from the probability distribution and Z denotes the final result of the randomization, then we have:

$$Z = X + Y$$

$$X = Z - Y$$

Now, the n instantiations of the probability distribution are known. The Y distribution is publically known. For large enough number of values of n, the distribution Z can be approximated closely by using different methods. By subtracting Y from Z it is possible to approximate the original distribution of X. This method is way better than the other methods. The classical sequential method is not good enough to find the original distribution form the distorted distribution of the data records. This is the strong example of the additive perturbation to achieve privacy preserving data mining.

At the end of the process, we only have a distribution containing behavior of X. Single records are hard to find. Furthermore the transformation is available along the single records. So the new data mining algorithms need to be designed to work with single variation rather than individual data records.

1.3.1 Advantages of Randomization

One of the key advantages of the randomization method is that in comparison to other methods it is simple. It does not require any information about the other records; it can work on individual records in a simple way.

In comparison to the k-anonymity which require knowledge about other records to work on the individual records. So, the process of randomization can be implemented at the data collection level.

It does not require the trusted server containing all the original records. It is very good strength of it.

1.3.2 Disadvantages of Randomization

There are some disadvantages to randomization method as it treats each record individually irrespective of their locality.

The records which are outliers are most likely to be identified than those which lie in between dense data records. During the attack these records will fall prey often than other records.

To overcome the problem, more complicated noise is added to transform each and every record which in turns increases the overhead to randomization. The adversarial attacks on the process of randomization include the background attack.

1.3.3 Applications of Randomization

Various data mining solutions use randomization method (Liu & Thuraisingham, 2006). Classification method uses the approach of randomization method. The perturbed transactions which are result of randomization are used in aggregate association rule mining. The extended version of the method has application OLAP and various other data mining methods. The distributed privacy preserving data mining uses multiplicative perturbation method for randomization.

1.4 Motivation

While mining the data the explicit identifiers are obviously removed. But still the identity can be revealed using quasi identifiers. The data from the two departments can help identify the individual with much accuracy. The records from medical department and the data from election department can be successfully matched to identify the individual. So there is need for privacy preservation in data mining. These quasi identifiers need to be secured or altered. The transformation methods such as perturbation provide much scope to explore the possible ways to enhance the privacy so that the sensitive data remains protected.

1.5 Objective of the work

There is different privacy objectives when anonymizing a dataset, these objectives are:

- **Unique identity disclosure**: – If data is published there must be no record that could identify an individual person.
- **Sensitive attributes disclosure**: – Adversaries cannot learn any sensitive information about any individual in the dataset via the disclosed attributes.

The main trade off is between privacy violation and information loss. The issue is addressed by combining the two techniques. First we apply randomization and then k-anonymity. In this work, the original data is randomized and then the sensitive attributes are classified into two categories high sensitive and low sensitive. The high sensitive attribute is transformed using k-anonymization and the low sensitive attribute is kept as it is.

1.6 Organization of the Thesis

This thesis is organized as follows. The chapter on introduction is followed by chapter 2 will be dedicated to literature review. It is followed by study on privacy preserving data mining techniques as chapter 3. In this chapter, comparison of advantages and disadvantages of the existing techniques is done. Chapter 4 explains and discusses the proposed approach. Chapter 5 is dedicated to implementation and analysis. The implementation is performed and it is tested on the synthetic data set. Chapter 6 concludes the report and discusses the future scope in the work done

CHAPTER 2

LITERATURE SURVEY

Following is the research work done on data perturbation using Randomization.

- The batch generation method is proposed by the author (Ellakkiya & Velvizi, 2013). In this method the data records are divided into batches. The data is requested by the data miner for various purposes. The miner request is then checked for the trust level. Based on the trust level the data miner is given either high perturbed copy of data or low perturbed copy of data. This method is called Multi Level privacy preserving data mining. In multilevel trust PPDM the sensitive attribute is distorted and protected by modifying it.
- In addition to additive and multiplicative perturbation techniques, there are other techniques which perturb the data using probability distribution. The techniques such as normal distribution, Gaussian distribution comes under probability distribution. The author (Chidambaram & Srinivasgan, 2014) proposes a combined random noise generation approach with multilevel privacy preservation data mining. In this approach, normal distribution is applied throughout the data set. Then Gaussian distribution is applied on sensitive attribute. And multilevel trust PPDM is applied to get maximum privacy to the sensitive data.
- The existing process of randomization is enhanced (Sharma et al.., 2014). In this method the attribute values from data set are perturbed. A probability matrix M is used to associate with every quasi identifier attribute of the data set. The method uses probability matrix for each quasi identifier. The technique is discussed in detail in chapter 3.

Following is the related work done in protecting privacy using k-anonymity:

- The author (Bettini, Wang, & Jajodia, 2006) uses method of k-anonymity in databases with time stamped information. It defines the k-anonymous views of temporal data. The research work also includes the extended approach on the generalization method so that it can fit to temporal data. The author points out the short coming of the generalization method of anonymization as it is difficult to apply on the time stamped data. The generalization algorithm is proposed mainly for time dependent data so that the time granularity hierarchy will be considered. Quasi identifiers are anonymized with the temporal data such that the data will be k-anonymous.

- The author (Zhu & Peng, 2007) analyzes the present condition of China medical information sharing. The paper exemplifies the generalization and suppression method of k-anonymity. He proposes the linking attack problem and possible solutions to it. The proposed method is compared with the classical algorithms and pros and cons are noted.
- The author (Blosser & Zhan, 2007) identifies the flaws in applying the k-anonymity to real time data in the medical organization and proposes the solution that releases the data with decreased privacy threat to the patient identity.
- The author (Zacharauli, Gkoulalas-Divanis, & Verykios, 2007) introduces the model of k-anonymity for spatial and temporal data. The Location Based Services are the cause for disclosure of individual information to large set of organizations. A generalization and unlinking algorithms are proposed to diminish the linkage of sensitive information. The unlinking algorithm severe the link between previous and the current request. A new form of quasi identifier is introduced here called as Location Based Quasi Identifier. It is a spatial-temporal movement pattern which consists of a series of spatial-temporal elements and recurrence formula. A physical history of locations is made to capture the spatial movement of the person. With the help of PHL and a series of requests by user, it is possible to identify the user. The unlinking algorithm is proposed as a solution for this problem.
- The author (Wu, Sun, & Wang, 2009) proposes a privacy preserving k-anonymity model for re-publication of incremental datasets. The paper states a monotonic generalization algorithm for the incremental data in reality to effectively prevent the privacy breach of the personal data. The proposed model makes it possible to be used in re-publication of the dynamic dataset. The concept of multi-dimensional partitioning is used here. The proposed method of **incremental data generalization** is used. In which it shown that how a new data is inserted in the database and it is generalized

dynamically. This generalization is further divided into group keeping generalization and group splitting generalization.

- The author (Shen, Liu, & Zhang, 2007) introduces new anonymity method called personalized granular k-anonymity. It pointed out the short comings of k-anonymity, saying that the parameters used are the system ones, constraints are overall situation based. They do not think of personalized demands, and local optimization. The method assigns the privacy degree to the attributes based on the individual level. For example, one person thinks fever is not a sensitive illness but the second person thinks that it is a sensitive illness; another one does not care about it.
- The author (Gong, Sun, & Xie, 2010) extended the work in protecting privacy in Location Based Services and proposed a method of k-anonymity without cloaked region. It says that the traditional k-anonymity method needs complex query processing at the server side. In this paper, author guarantee that it can make the record r anonymous among k records with low querying cost and communication cost. With comparison of traditional algorithm for k-anonymity which uses complex algorithms at server side, the proposed method uses only in query processing algorithm.
- The author (Vijayrani, Tamilarasi, & Sampoorna, 2010) analyses the privacy preserving k-anonymity methods and techniques. The method of k-anonymity is successful only if the dataset satisfies it requirement that it should have at least k occurrences of the record to apply. The paper classifies the k-anonymity method into two types as the data reduction approach and the data perturbation approach. United States people are subject to privacy threat as they can be uniquely identified by combination of attributes like age, gender, date of birth, zip code, etc. Various medical data in the organizations are supposed to be anonymous but it can be uniquely identified using linking attack. The paper states that the method of k-anonymity has two assumptions. The first assumption is that the data owner knows about the sensitivity of the data. That is he is able to separate the data set attributes into quasi identifiers and private attributes. The second assumption is that the attacker has full access to the public parameters about the individual and don’t have access to the private information.

- The author (Wang et al.., 2011) applied the k-anonymity on the transactional databases. The data from the transactional dataset is classified into two type sensitive items and quasi identifiers. The non-sensitive attribute are let at it is. Anonymization is performed on the sensitive items from the transactional data. The paper presented two algorithms Sensitive Transaction Neighbors and Gray Sort Clustering to achieve sensitive k-anonymity. The first approach finds the sensitive traction with least sensitive items and then using hamming distance it finds other related transactions which will be equal to k-1. The process repeats itself so that all the sensitive items get exhausted. The basic idea of the second approach is to find the transactions with similar non sensitive items. It basically starts transaction permutations which searches same number of transactions above and below it.
- The author (Li & Liu, 2011) proposes a model for representation of k-anonymity worlds. An efficient representation that takes less space is presented for anonymous data set. The model also specifies the efficient query processing method for anonymous data.
- The author (Liu, Jia, & Han, 2012) presents an improvement to the k-anonymity algorithm to apply for the dataset which has multiple sensitive attributes. The proposed method sorts the sensitive attributes and tuples based on greedy strategy.

The extended k-anonymity work done can be summarized with following tables:

Table 2.1 Original data

Age	**Gender**	**ZIP**	**Income**	**Marital Status**	**Disease**
27	M	26022	10000	Single	Fever
45	M	29915	17000	Married	Cancer
32	M	21988	40000	Divorced	AIDS
34	M	29488	10000	Single	Fever

Here, the sensitive attributes are income, marital status, disease.

Table 2.2 4-anonymized data set

Age	Gender	ZIP	Income	Marital Status	Disease
[20-50]	M	2****	10000	Single	Fever
[20-50]	M	2****	17000	Married	Cancer
[20-50]	M	2****	40000	Divorced	AIDS
[20-50]	M	2****	10000	Single	Fever

Table 2.3 4-anonymized data set using new algorithm

Age	Gender	ZIP	Income	Marital Status	Disease
[20-50]	M	2****	10000(2) 17000 40000	Single(2) Married Divorced	Fever(2) Cancer AIDS
[20-50]	M	2****			
[20-50]	M	2****			
[20-50]	M	2****			

The other work done on the anonymization method includes:

- The author (Burke & Kayem, 2014) applies the method of k-anonymity on the crime data where the resource environment is constrained. Many times the person won't report the crime incidence fearing to the privacy. Crime data mining is an emerging field where the privacy preservation should be applied. Here the less offensive crimes which leads to much less punishment compared to the heinous crime are made anonymous.

- Generalization is often used in combination with suppression. As generalization causes more data degradation, suppression is performed on specific tuples where there is less number of outliers. The notion of k-anonymity is proposed by (Samarati, 2001). In this approach she was using k-anonymity for de-identification of the microdata. The governmental, public, private institution collect data which in tabular form. This electronic data is called microdata.

The author (Basu et al.., 2015) evaluates the risks and reality with k-anonymity method. The problem of optimal k-anonymization is NP-hard. The probability of re-identifying the anonymized data is termed as risk. The possible methods through which the k-anonymity is applied are analyzed and risk is calculated on the different parameters.

Following is the research work done with l-diversity as the keyword:

- The author (Jian-min, Ting-ting, & Hui-qun, 2008) introduces the notion of micro aggregation. An aggregator operator (for example the mean for continuous data or mode for categorical data) is computed for each cluster for each cluster's centroid which is used to replace the original record. In other words each record in a cluster is replaced by the cluster's centroid. There is another notion of k-partition. The records are divided into partition sot that the most similar records fall into same partition and the size of the partition should not exceeds k. The optimal micro aggregation is possible when there is optimal k-partition, which requires maximizing the homogeneity within the group. The homogeneity of the group is dependent on various other factors like the kind of data, and knowledge to be extracted. The author proposes a micro aggregation algorithm for l-diversity to apply it on numerical data. The paper also defines the kinds of l-diversity. The types of l-diversity are made under the notion of distinct diversity degree, entropy diversity degree, and diversity degree of table. The proposed algorithm does not consider sensitive attribute diversity of the equivalence class.
- The author (Han, Yu, & Yu, 2008) also proposes an improved approach for l-diversity of numerical data through l-incognito algorithm. L-incognito algorithm starts with checking l-diversity of single attribute subsets of the quasi identifier

and a graph of single attribute generalization is constructed. Then the above process is iterated with the increase in size of the subsets, till the size of the subsets equal the quasi identifiers. On every iteration a new graph of multi attribute generalization is constructed based on the graph constructed at previous level. And at each process of graph generation it prunes the generalizations which cannot satisfy l-diversity. It is based on the theory that says if there is a set of attributes P such that P is subset of Q such that T does not satisfies l-diversity with respect to P, then T does not satisfy the l-diversity requirements with respect to Q.

L-incognito algorithm: It is a high efficient domain generalization k-anonymization algorithm. Since, it could not implement the diversity of sensitive attributes, the anonymized table generated by the algorithm fall to homogeneity attack and background knowledge attack. This algorithm is designed based on Incognito algorithm. Domain generalization and value generalization are the two ways in which the quasi identifiers are generalized. Domain generalization is easy and simple to apply but the information loss is high. On the other hand value generalization is complicated to apply but it retains the information.

If the dataset has the skewed distribution of sensitive attributes then there is probability of losing data or privacy due to existing methods (Tian & Zhang, 2009). The problem is solved by generalizing the sensitive attributes in the data sets; also a function is proposed to constrain the frequencies of sensitive attribute values.

CHAPTER 3

PRIVACY PRESERVING DATA MINING TECHNIQUES

3.1 Matrix method for Randomization

In this method the attribute values from data set are perturbed (Sharma et al.., 2014). A probability matrix M is used to associate with every quasi identifier attribute of the data set. Following are the steps involved:

- The existing method associates or links each value from attribute to particular element of the matrix.
- The particular element is found with searching algorithm.
- The highest element from the row is selected for corresponding attribute value. That is column value is associated with highest row value from the matrix.
- If there are more than one value which is highest then the left most one from the row is selected for assignment.
- Generally the size of the matrix is equal to the size of column. And it is a square matrix.
- As the associated elements are sorted so as the column values.

3.2 Anonymization

The method anonymization is also known as k-anonymity. It works such that a particular record would not be exposed explicitly. It is termed as principle of k-anonymity. The records are made similar so that it is difficult to identify specific record. Following tables will show the example of anonymization.

Table 3.1 Original Data

Name	Age	Zip code
Ram	43	444232
Shyam	26	444123
Jim	56	444124
Mary	23	444125

Table 3.2 Anonymized data

Name	Age	Zip code
*	43	444232
*	20-30	4441**
*	50-60	4441**
*	20-30	4441**

The anonymization can be implemented using grouping the records, using range instead of using particular value, substituting another value for representation. For example, the attribute value M/F can be replaced with 0/1 in numeric.

The value of ‘k’ in k-anonymity varies to the size of group. In above example value of k is 2 with respect to the attribute age and zip code.

The other example of k-anonymity is:

Table 3.3 3-anonymized data set

Age	Weight	Height	Disease
[20-40]	[50-70]	[5-7]	Blood Cancer
[30-50]	[60-70]	[5-7]	AIDS
[30-50]	[60-70]	[5-7]	Mental Illness
[30-50]	[60-70]	[5-7]	Cancer
[30-50]	[60-70]	[5-7]	Parkinson’s

The above two methods are known as suppression and generalization.

3.2.1 Suppression: In this method the attribute value is replaced by asterisk (*). Either the whole value is replaced or certain part of the value is replaced by *. In above example attribute name is suppressed.

Suppression can be applied on the following levels in the dataset:

a. Tuple: whole tuple can be removed from the dataset. One should be careful to apply suppression as it radically degrades the quality of data.

b. Attribute: suppression is applied on whole column. Whole column is removed from the table.

c. Cell: suppression is done on a single cell. A particular cell from tuple or from the column is removed.

3.2.2 Generalization: In this method the attribute value is replaced by its range. For example age 23 is replaced by attribute value ‘20-30’. Or it can also be replaced by ‘age<30’.

Generalization is performed on different levels:

a. Attribute: generalization is done on the column; whole column values are replaced to a general value.

b. **Cell:** a particular cell value is generalized from the column. A cell from the specific attribute is modified to a general value. For example, an instance in age column.

In many organizations the data records are made available by simply removing the explicit identifiers such as name, key identifiers, and social security numbers. But the quasi identifiers are let at it is. The combination of these attribute can be significant to reveal the real identity.

The approach of k-anonymity assumes an ordering among the quasi identifier attributes. The values in the tuples are discretized into intervals or grouped into different number of values. The categorization is performed the discretized values.

3.2.3 Attacks on k-anonymity

The method of k-anonymity suffers from basically two attacks (Machanavajjhala et al.., 2006).

- **Homogeneity attack:** The dataset which contains identical values of the attributes suffers from homogeneity attack. In this case even if the k-anonymity is performed on the data set, the original data can be revealed.
- **Background knowledge attack:** The data set which belongs to the particular community in the society can be fall threat to back ground knowledge attack. The anonymized quasi identifiers which are associated with the sensitive attribute can be reduced to get original values or approximate values.

The two attacks can be explained by using an example. Consider the following data set.

Table 3.4 Medical data set

ZIP	AGE	DISEASE
444321	29	Heart Disease
444231	22	Heart Disease
444132	27	Heart Disease
443321	43	Flu
445223	52	Heart Disease
445224	47	Cancer
445234	30	Heart Disease
443123	36	Cancer
443233	32	Cancer

Table 3.5 3-anonymized version of medical dataset

ZIP	AGE	DISEASE
444***	2*	Heart Disease
444***	2*	Heart Disease
444***	2*	Heart Disease
4452**	>40	Cancer
4452**	>40	Flu
4452**	>40	Heart Disease
443***	3*	Heart Disease
443***	3*	Cancer
443***	3*	Cancer

The Table 3.4 is the original data set form the medical institute and the Table 3.5 is the 3-anonymous version of the data. Bob is living at ZIP code 444321 and has Heart disease. Alice has the information about where Bob lives. Then she can easily know that Bob has Heart Disease from homogeneity attack. As all the records corresponding to first equivalence class have identical values. Suppose, Alice knows age and the ZIP code of the place where Bob lives, and by observing the table it concluded the record must be from the last class. Also she knows that Bob is having very little chance of Heart Disease, Alice revealed that Bod has cancer. This is an example of background knowledge attack.

There are other techniques which are developed to overcome the limitations of k-anonymity. L-diversity and t-closeness are the extension of k-anonymity.

3.3 Cryptographic Techniques

Cryptography involves mathematical and logical transformation of the data. The various encryption methods used to secure the data can be used for privacy preserving data mining. (Abitha et al.., 2015) proposes a cryptographic approach to privacy preserving data mining. The rail fence algorithm and Vigenere cipher algorithms are modified to use in this approach. The original data is itself used to create the matrix in Vigenere cipher algorithm. With the experiments it is proved that the original data cannot be inferred from the modified data.

3.3.1 A Rail fence algorithm

This technique is mostly applied to the categorical data where data is written in row/column wise and it is retrieved by traversing along column/row wise. For example, following text is written to transform it: THIS TEXT IS RAIL FENSED.

T	E	S	I	
	H	X	R	L
		I	T	A
			S	I
				T

Figure 3.1 a rail fence algorithm

So the text will be written as: TESI HXRL ITA SI T.

In above approach no key is used so the encryption overhead using the key is reduced. The author has modified the algorithm for applying it for the purpose data perturbation.

3.3.2 Vigenere cipher algorithm

This algorithm is used mainly for categorical data. The noise generated in the algorithm will contain the group of characters. The intermediate matrix will contain letter from A to Z. A specific key is fixed. The key is applied to transform the data values.

3.3.3 Disadvantages of Cryptographic technique

From both the algorithms, it is clearly visible that the data is encrypted. And if the attacker knows about the algorithm then it just matter of time that it will find the key and the original data. Also the overhead of encryption increases as the size of data increases. There are data sets which have billions trillions of values and hundreds of attributes. In this case the applying data encryption techniques prove to be disadvantage as it increases the data size. Also the approach consumes more resources.

3.3.4 Advantages of cryptographic techniques

These techniques are feasible to apply on small scale of data. The data utility after the transformation remains intact as the original data can be recovered by reversing the technique. These techniques can be applied where there is trusted connection between the organizations.

3.4 A soft computing model for PPDM

The soft computing model using neural network and fuzzy logic is presented by (Malik, Asger, & Sarvar, 2015). Neural networks have been proven to be cost effective solution to various problems in data mining. So it will prove to be an efficient solution for privacy preserving data mining. Soft computing methods have emerged as a powerful tool as it is tolerant to uncertainty, partial truth, and imprecision. It helps in achieving solutions that are having low cost, robust, and practically implementable. Health care systems use neural networks extensively for analysis purposes in every field from supplying to distribution.

3.4.1 Soft Computing

The problems of data mining field are well solved by soft computing methodologies. Soft computing is different from conventional hard computing. Soft computing works like putting human mind as role model. Machine Intelligence, Fuzzy logic, rough computing, neural networks, Genetic algorithms are among the most commonly used methodologies in soft computing. Uncertainty is very effectively dealt by Fuzzy set. Neural network is composed of very large number interconnected neurons. As per our brain has neurons so the architecture which is inspired by it uses them in the technology.

Neural networks are successfully implemented in rule generation and optimization. Search and optimization algorithms use genetic algorithms. Rough set is used to manage uncertainty. Soft computing is different and unique methodology that work in as robust and cost effective manner so that it is always feasible to apply to the problem. The intelligent and robust system developed is comparable to the conventional techniques.

3.4.2 Fuzzy-neural Model

This model works as follow. The data from various sources is collected. It removes the explicit and the quasi identifiers from the data set and then fuzzify the sensitive attribute using fuzzy membership function. The model is then tested for the result. The efficiency of the model is checked using confusion matrix. The relationship between the actual and predicted classification is represented using confusion matrix. The confusion matrix shows that there is not much difference between the actual and the transformed data. So the information loss is reduced.

By the above work it is shown that soft computing can prove to be a good technique in preserving privacy of an individual. But the data utility is diminished in the process of fuzzyfying the data set. Since all the quasi identifiers are also removed.

3.5 l-diversity

L-diversity is an extension to k-anonymity. K-anonymity has shortcomings which lead to various attacks on it. The homogeneity attack act when the attribute values are identical. L-diversity method decreases the granularity of the dataset representation. (Machanavajjhala et al.., 2007) introduced the strong notion of l-diversity addressing the shortcomings of the k-anonymity. The principle of l-diversity is stated as:

Principle of l-diversity: An equivalence class is said to be in l-diversity if at least there are l ‘well represented’ sensitive attributes.

The paper also defines various other notions to extend the principle of l-diversity. The discussions performed in the paper on the topic are:

- L-diversity does not require knowledge of full distribution of sensitive and non-sensitive attributes.
- There is no need of as much information to the data publisher as it require for the adversary. The parameter l keeps the data secure against the most knowledgeable adversaries. As the value of l increases then the amount of information is also increases to rule out the possible values of sensitive attributes.
- Instance knowledge is covered up automatically. Suppose, Alice knows that there are no chances of Bob having heart disease. This knowledge is suppressed by l-diversity. It is the same way as treating the possible values of sensitive attributes.
- Different inferences are caused due to the different adversaries having different kind of knowledge. L-diversity secures the data without the need of having information about the level of background knowledge the adversary has.

3.6 t-closeness

This approach to k-anonymity is further enhancement of l-diversity. It is defined as ‘An equivalence class is said to have t-closeness if the distance between the distribution of a sensitive attribute in this class and the distribution of the attribute in the whole table is no more than a threshold t. A table is said to have t-closeness if all equivalence classes have t-closeness’.

Intuitively, privacy is measured by the information gain of an observer. Before seeing the released table, the observer has some prior belief about the sensitive attribute value of an individual. After seeing the released table, the observer has a posterior belief. Information gain can be represented as the difference between the posterior belief and the prior belief (Li, Li, & venkatsubramanian, 2007).

One characteristic of the *l*-diversity model is that it treats all values of a given attribute in a similar way irrespective of its distribution in the data. This is rarely the case for real data sets, since the attribute values may be much skewed. This may make it more difficult to create feasible *l*-diverse representations. Often, an adversary may use background knowledge of the global distribution in order to make inferences about sensitive values in the data. Furthermore, not all values of an attribute are equally sensitive. For example, an attribute corresponding to a disease may be more sensitive when the value is positive, rather than when it is negative.

3.7 Distributed privacy preserving data mining

The key goal in most distributed methods for privacy-preserving data mining 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.

The problem of distributed privacy-preserving data mining overlaps closely with a field in cryptography for determining secure multi-party computations. A broad overview of the intersection between the fields of cryptography and privacy-preserving data mining may be found in (Pinkas, 2002). The broad approach to cryptographic methods tends to compute functions over inputs provided by multiple recipients without actually sharing the inputs with one another. For example, in a 2-party setting, Alice and Bob may have two inputs x and y respectively, and may wish to both compute the function $f(x, y)$ without revealing x or y to each other. This problem can also be generalized across k parties by designing the k argument function $h(x1 \ . \ . \ . \ xk)$. Many data mining algorithms may be viewed in the context of repetitive computations of many such primitive functions such as the scalar dot product, secure sum etc. In order to compute the function $f(x, y)$ or $h(x1 \ . \ . \ . \ , xk)$, a *protocol* will have to designed for exchanging information in such a way that the function is computed without compromising privacy. We note that the robustness of the protocol depends upon the level of trust one is willing to place on the two participants Alice and Bob. This is because the protocol may be subjected to various kinds of adversarial behavior:

- **Semi-honest Adversaries:** In this case, the participants Alice and Bob are curious and attempt to learn from the information received by them during the protocol, but do not deviate from the protocol themselves. In many situations, this may be considered a realistic model of adversarial behavior.
- **Malicious Adversaries:** In this case, Alice and Bob may vary from the protocol, and may send sophisticated inputs to one another to learn from the information received from each other.

A key building-block for many kinds of secure function evaluations is the 1 out of 2 oblivious-transfer protocol. This protocol was proposed and involves two parties: a *sender*, and a *receiver*. The sender's input is a pair ($x0$, $x1$), and the receiver's input is a bit value $\sigma \in \{0, 1\}$. At the end of the process, the receiver learns $x\sigma$ only, and the sender learns nothing. A number of simple solutions can be designed for this task. In one solution, the receiver generates two random public keys, $K0$ and $K1$, but the receiver knows only the decryption key for $K\sigma$. The receiver sends these keys to the sender, who encrypts $x0$ with $K0$, $x1$ with $K1$, and sends the encrypted data back to the receiver. At this point, the receiver can only decrypt $x\sigma$, since this is the only input for which they have the decryption key. We note that this is a semi honest solution, since the intermediate steps require an assumption of trust. For example, it is assumed that when the receiver sends two keys to the sender, they indeed know the decryption key to only one of them. In order to deal with the case of malicious adversaries, one must ensure that the sender chooses the public keys according to the protocol. An efficient method for doing so is described in (Naor & Pinkas, 2001). In this, generalizations of the 1 out of 2 oblivious transfer protocols to the 1 out *N* case and *k* out of *N* case are described.

Since the oblivious transfer protocol is used as a building block for secure multi-party computation, it may be repeated many times over a given function evaluation. Therefore, the computational effectiveness of the approach is important. Efficient methods for both semi-honest and malicious adversaries are discussed in (Naor & Pinkas, 2001). More complex problems in this domain include the computation of probabilistic functions over a number of multi-party inputs. Such powerful techniques can be used in order to abstract out the primitives from a number of computationally intensive data mining problems. Many of the above techniques have been described for the 2-party case, though generic solutions also exist for the multiparty case. Some important solutions for the multiparty case may be found in (Chaum, Crepeau, & Damgard , 1988).

The oblivious transfer protocol can be used in order to compute several data mining primitives related to vector distances in multi-dimensional space. A classic problem which is often used as a primitive for many other problems is that of computing the scalar dot-product in a distributed environment. A fairly general set of methods in this direction are described in (Du & Atallah, 2001). Many of these techniques work by sending changed or encrypted versions of the inputs to one another in order to compute the function with the different alternative versions followed by an oblivious transfer protocol to retrieve the correct value of the final output. A systematic framework is described to transform normal data mining problems to secure multi-party computation problems. The problems discussed in include those of clustering, classification, association rule mining, data summarization, and generalization. A second set of methods for distributed privacy-preserving data mining is discussed in (Clifton et al.., 2002) in which the secure multi-party computation of a number of important data mining primitives is discussed. These methods include the secure sum, the secure set union, the secure size of set intersection and the scalar product. These techniques can be used as data mining primitives for secure multi-party computation over a variety of horizontally and vertically partitioned data sets. Next, we will discuss algorithms for secure multi-party computation over horizontally partitioned data sets.

CHAPTER 4

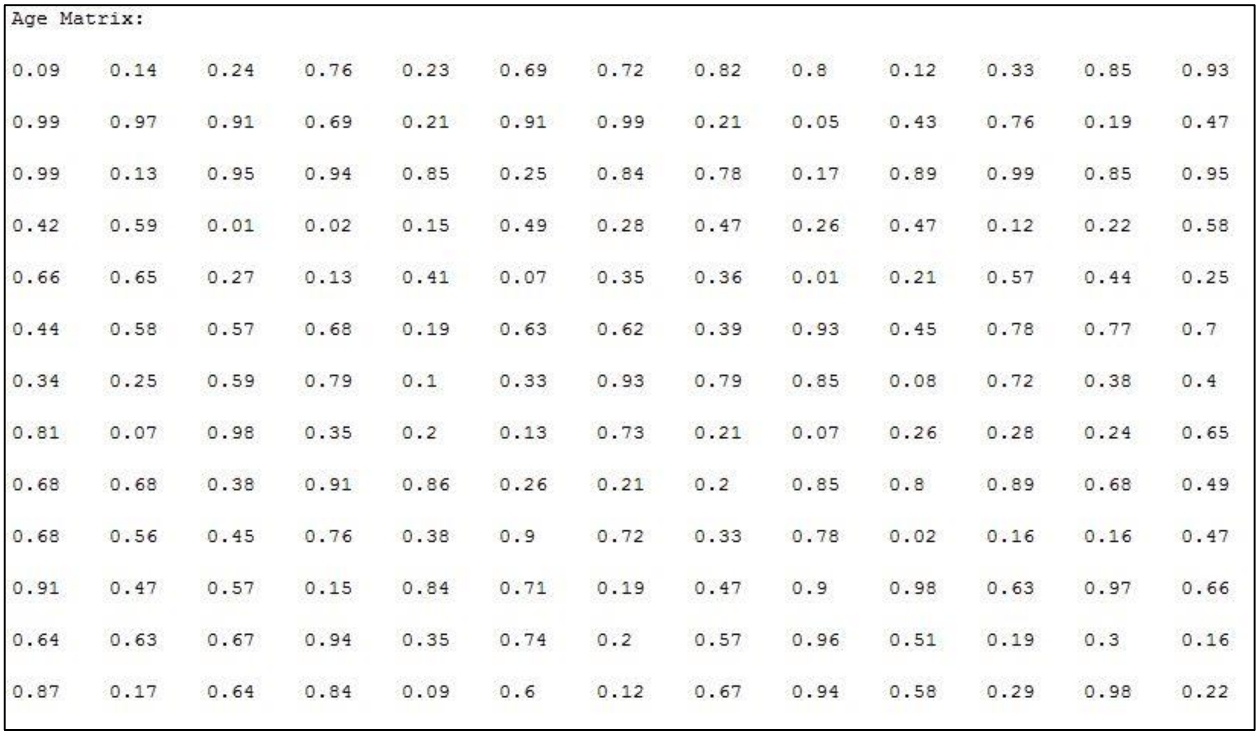
PROPOSED APPROACH

The randomization and the anonymization techniques are used in combination to get an efficient approach for privacy preserving data mining. In randomization process, data perturbation using matrix method is implemented.

4.1 Proposed Matrix Method:

- The matrix is created randomly same as the procedure discussed in chapter 3.
- Then the matrix size is reduced from square matrix to some fixed constant in our case it is 5. Hence the new size of the matrix becomes [size of col, 5].
- Then the matrix is traversed diagonally from (1, 1) to (5, 5). Then it is reverse tracked from (5, 5) to _ and so on.
- While traversing the elements are mapped to sequentially to the column elements.
- Then the elements are sorted and with it the column values also get sorted. If there are two values then they are kept one after another.
- This is how the quasi identifiers get perturbed using matrix method.

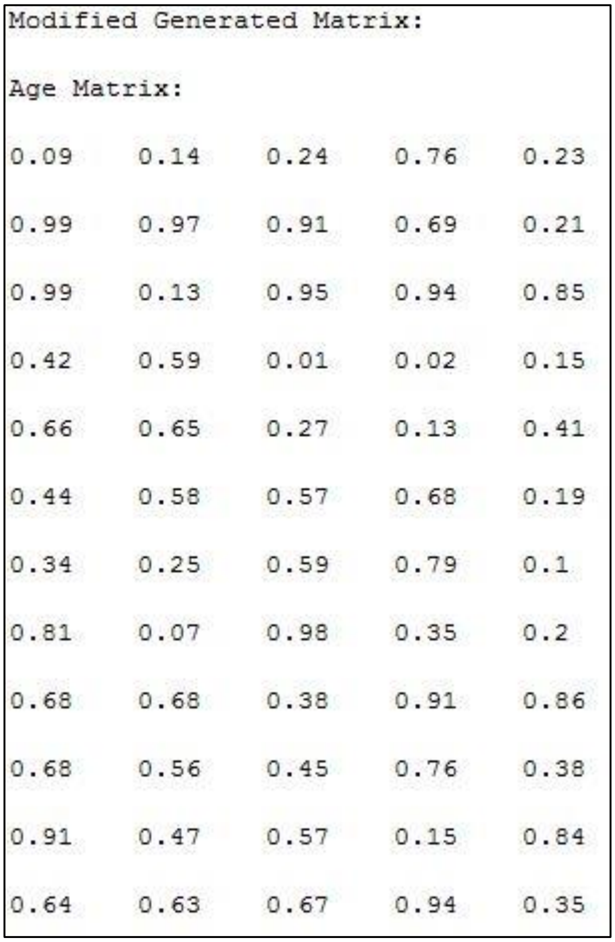
The following figure shows the example of the matrix:



```
Age Matrix:

0.09    0.14    0.24    0.76    0.23    0.69    0.72    0.82    0.8     0.12    0.33    0.85    0.93

0.99    0.97    0.91    0.69    0.21    0.91    0.99    0.21    0.05    0.43    0.76    0.19    0.47

0.99    0.13    0.95    0.94    0.85    0.25    0.84    0.78    0.17    0.89    0.99    0.85    0.95

0.42    0.59    0.01    0.02    0.15    0.49    0.28    0.47    0.26    0.47    0.12    0.22    0.58

0.66    0.65    0.27    0.13    0.41    0.07    0.35    0.36    0.01    0.21    0.57    0.44    0.25

0.44    0.58    0.57    0.68    0.19    0.63    0.62    0.39    0.93    0.45    0.78    0.77    0.7

0.34    0.25    0.59    0.79    0.1     0.33    0.93    0.79    0.85    0.08    0.72    0.38    0.4

0.81    0.07    0.98    0.35    0.2     0.13    0.73    0.21    0.07    0.26    0.28    0.24    0.65

0.68    0.68    0.38    0.91    0.86    0.26    0.21    0.2     0.85    0.8     0.89    0.68    0.49

0.68    0.56    0.45    0.76    0.38    0.9     0.72    0.33    0.78    0.02    0.16    0.16    0.47

0.91    0.47    0.57    0.15    0.84    0.71    0.19    0.47    0.9     0.98    0.63    0.97    0.66

0.64    0.63    0.67    0.94    0.35    0.74    0.2     0.57    0.96    0.51    0.19    0.3     0.16

0.87    0.17    0.64    0.84    0.09    0.6     0.12    0.67    0.94    0.58    0.29    0.98    0.22
```

Figure 4.1 Generated age matrix

The generated matrix is modified which is shown below:



```
Modified Generated Matrix:

Age Matrix:

0.09    0.14    0.24    0.76    0.23

0.99    0.97    0.91    0.69    0.21

0.99    0.13    0.95    0.94    0.85

0.42    0.59    0.01    0.02    0.15

0.66    0.65    0.27    0.13    0.41

0.44    0.58    0.57    0.68    0.19

0.34    0.25    0.59    0.79    0.1

0.81    0.07    0.98    0.35    0.2

0.68    0.68    0.38    0.91    0.86

0.68    0.56    0.45    0.76    0.38

0.91    0.47    0.57    0.15    0.84

0.64    0.63    0.67    0.94    0.35
```

Figure 4.2 Modified age matrix

4.2 Comparison Between the two approaches

The existing method applies searching and sorting on the matrix that is being mapped to the attribute for randomization. The algorithm for the existing approach can be as follow:

1. Scan the table for key attributes.
2. Remove the key attributes.
3. Now, the table is left with quasi identifiers (attributes which are less sensitive) and sensitive attributes.
4. Each quasi identifier is mapped to the element of the matrix.
5. The matrix is created randomly. The size of the matrix is equal to the size of the quasi identifier.

6. From every row largest element is selected and it is mapped to the element from the attribute.
7. If two largest elements are equal then left element is selected for mapping.
8. As the matrix elements are sorted so the corresponding identifiers are also sorted.
9. And thus the quasi identifiers get randomized.
10. Generalization is performed on sensitive attribute.

The method which is presented can be improved. The matrix elements used for randomization are firstly searched and then sorted. So it consume more time. Again the in anonymization the quasi identifier gender is kept as it is. The change is made to it and it is replaced by 0/1 as it will be helpful to extend the anonymization.

In the proposed approach enhancements are made. The proposed procedure is as follow.

1. Scan the table for key attributes.
2. Remove the key attributes.
3. Now, the table is left with quasi identifiers (attributes which are less sensitive) and sensitive attributes.
4. Each quasi identifier is mapped to the element of the matrix.
5. The matrix is created randomly. The size of the matrix is equal to the size of the quasi identifier.
6. The matrix is converted to non-square matrix. It is traversed diagonally. As it goes from one diagonal element to another diagonal element, these elements are mapped to corresponding identifier elements.
7. These matrix elements are then sorted.
8. If two elements are mapped to equal value then they are kept in original order as in table.
9. This is how quasi identifier gets randomized faster.
10. The sensitive attribute is then anonymized using generalization.

4.3 The generalization method

The generalization process explained in previous chapter under anonymization is used as the second method to be applied on the data set. The quasi identifier attributes which are randomized are used as the input for the process. The medical data set used for experiment contains quasi identifiers as age, gender, ZIP code. Age attribute is generalized to using range as the substitute to the single value. A range of 10 years is made for the replacement of the attribute value. Then gender attribute values are altered to 0/1 to substitute the binary attribute values (Kaur & Bansal, 2015). The next attribute ZIP code is generalized by adding special character like ’*’ at the end of the ZIP code value replacing the digits. The number of digits to be replaced depends on the attribute value. Then the data set is recombined.

4.4 Flow charts for the proposed method

The proposed approach is represented using flow chart diagram which will elaborate on the process more. Figure 4.3 states about randomization method and Figure 4.4 states about anonymization.

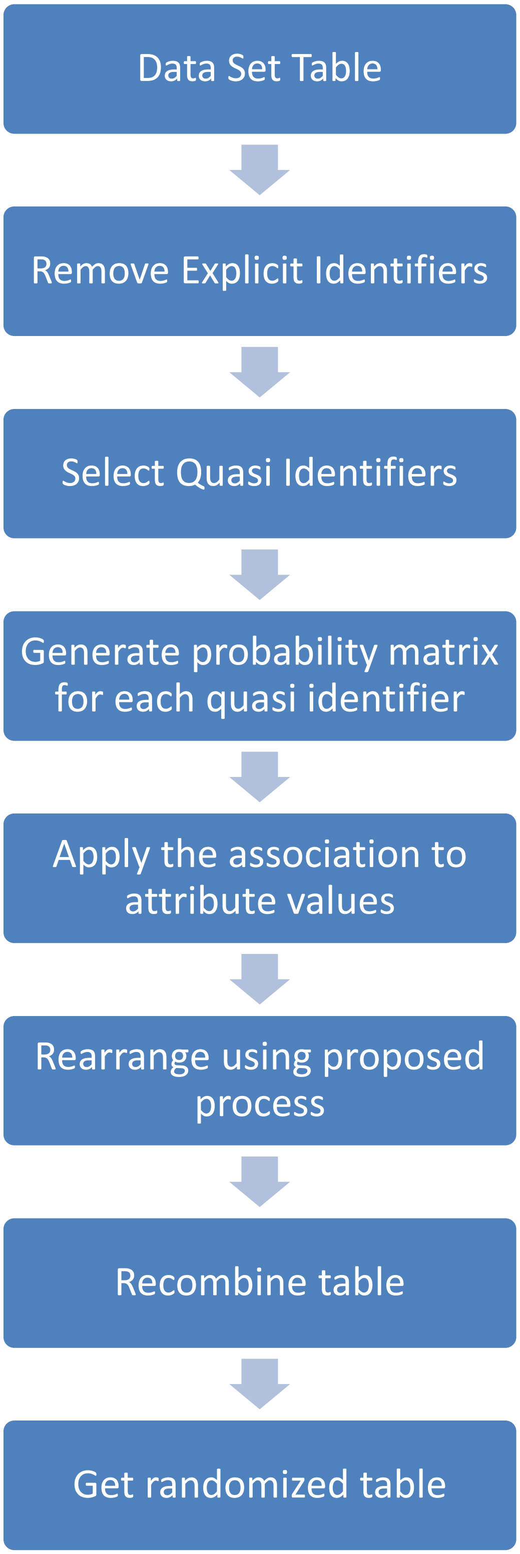

Data Set Table
Remove Explicit Identifiers
Select Quasi Identifiers
Generate probability matrix for each quasi identifier
Apply the association to attribute values
Rearrange using proposed process
Recombine table
Get randomized table

Figure 4.3: Flowcharts for Randomization

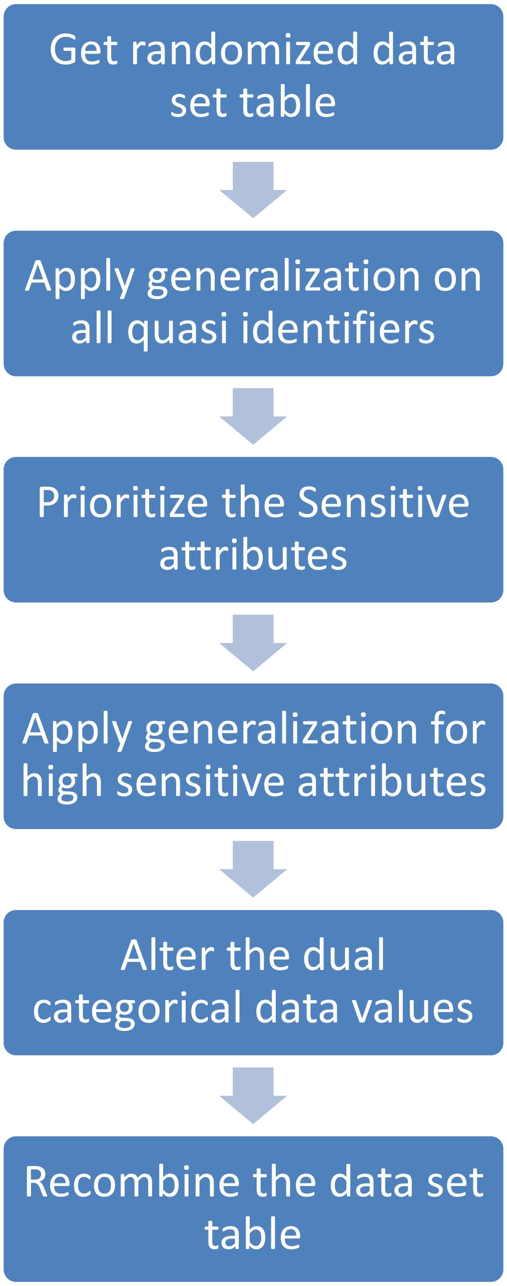

Get randomized data set table
Apply generalization on all quasi identifiers
Prioritize the Sensitive attributes
Apply generalization for high sensitive attributes
Alter the dual categorical data values
Recombine the data set table

Figure 4.4 Flowcharts for anonymization

Here in anonymization process changes are made to the existing approach too. The Unlike the existing approach, generalization is performed on all the quasi identifiers values. And grouping generalization is performed on the high sensitive attribute values only.

CHAPTER 5

IMPLEMENTATION AND ANALYSIS

5.1 Environmental Setup

The following conditions were used while implementation of the proposed approach.

- **Hardware configuration**

Processor	: Intel core i3
Processor speed	: 2.40 GHz
RAM storage	: 4 GB
Hard Disk Storage	: 500 GB

- **Software configuration**

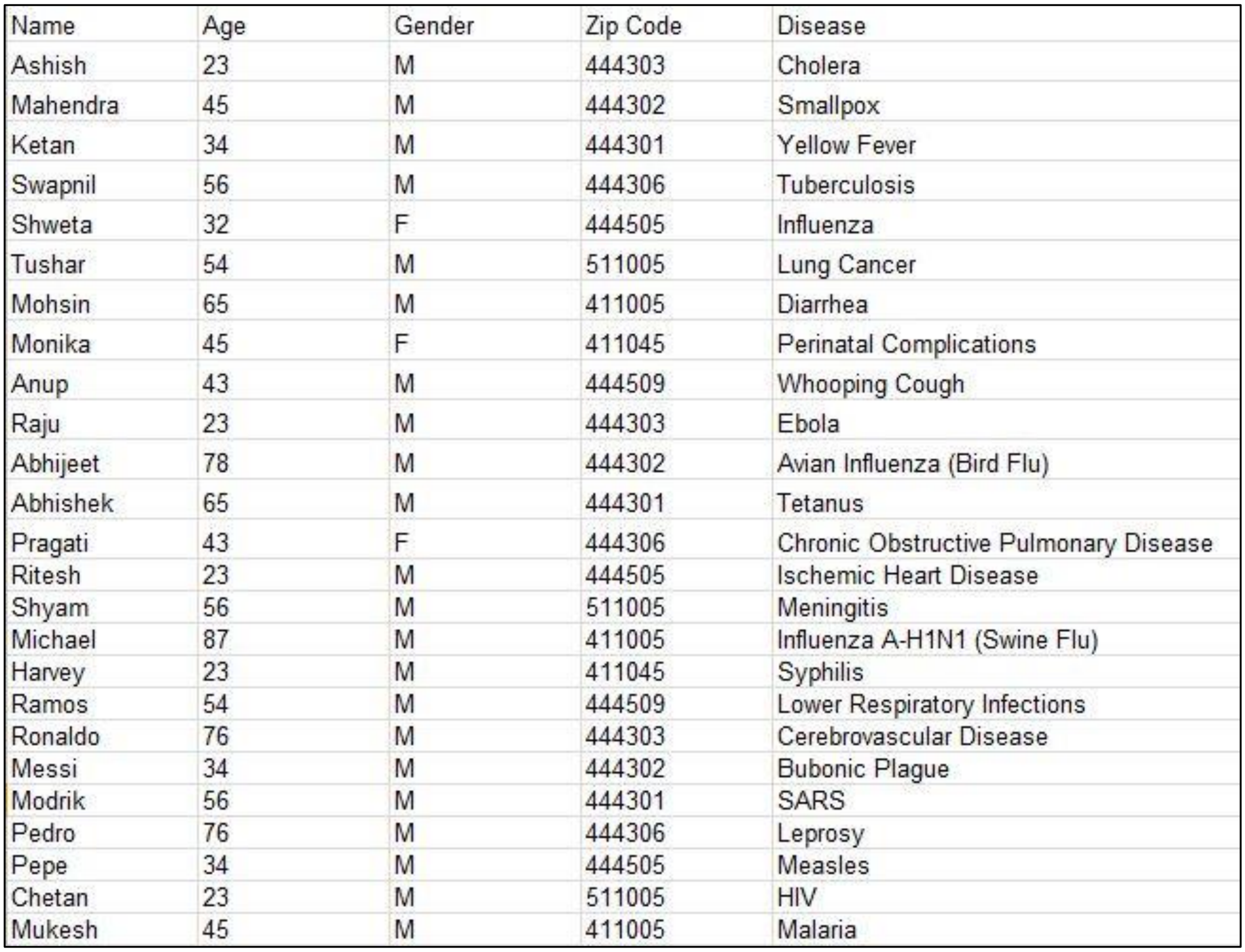
Operating system	: Windows 8.1 Pro
Language used	: Java
IDE used	: Spring Tool Suite
Data Storage	: Microsoft Excel

5.2 Result of application of proposed method

The proposed method is applied on medical data set. Here the attributes are name, age, gender, ZIP code, disease. These attributes can be classified as:

- Explicit Identifiers : name
- Quasi Identifiers : age, gender, ZIP code
- Sensitive attributes : disease

The proposed work will be explained by the running example of sample data set. The original data set is like in following figure.



Name	Age	Gender	Zip Code	Disease
Ashish	23	M	444303	Cholera
Mahendra	45	M	444302	Smallpox
Ketan	34	M	444301	Yellow Fever
Swapnil	56	M	444306	Tuberculosis
Shweta	32	F	444505	Influenza
Tushar	54	M	511005	Lung Cancer
Mohsin	65	M	411005	Diarrhea
Monika	45	F	411045	Perinatal Complications
Anup	43	M	444509	Whooping Cough
Raju	23	M	444303	Ebola
Abhijeet	78	M	444302	Avian Influenza (Bird Flu)
Abhishek	65	M	444301	Tetanus
Pragati	43	F	444306	Chronic Obstructive Pulmonary Disease
Ritesh	23	M	444505	Ischemic Heart Disease
Shyam	56	M	511005	Meningitis
Michael	87	M	411005	Influenza A-H1N1 (Swine Flu)
Harvey	23	M	411045	Syphilis
Ramos	54	M	444509	Lower Respiratory Infections
Ronaldo	76	M	444303	Cerebrovascular Disease
Messi	34	M	444302	Bubonic Plague
Modrik	56	M	444301	SARS
Pedro	76	M	444306	Leprosy
Pepe	34	M	444505	Measles
Chetan	23	M	511005	HIV
Mukesh	45	M	411005	Malaria

Figure 5.1 Original data set

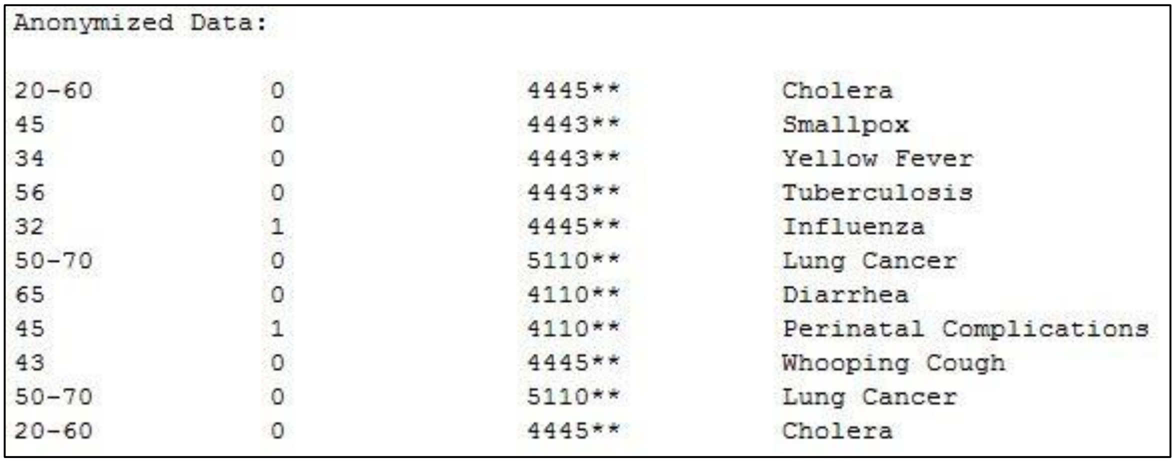
The original data set will be transformed to the following anonymized data set. Following steps are performed to get result.

- Explicit identifiers are removed. Here name is an explicit identifier; thus it is removed from the data set.
- First the randomization method is applied to the table. In which separate probability matrices are created for each quasi identifiers. The age, gender, and ZIP code attribute values are randomized using modified matrix method. The attribute values which are linked to diagonal values in way explained in the

algorithm are sorted according to an increasing order of the matrix elements. Thus the quasi identifier attributes get randomized.

- Now, the attribute ZIP code is generalized to four digits rather than six so that the area it refers enlarge. Normally six digit numbers refers to specific small part of the city or village. The four digit ZIP code will refer to the greater part of the city or the city itself. The last two digits of the ZIP code value are replaced by *.
- Next, the gender attribute has only two values that is M/F. These two values are altered with another binary set because the M/F values are too common in use. So there is possibility that it can lead to data discretization. Hence, these values are replaced by 0/1 to confuse the attacker.
- The age attribute is generalized forming a range. The range is taken as ten years. Again, only those age values are randomized which falls in line in which high sensitive disease value is present. The high sensitive group of disease includes AIDS, Cancer, Cholera, etc.

The final anonymized data is show in following figure:



```
Anonymized Data:

20-60          0              4445**         Cholera
45             0              4443**         Smallpox
34             0              4443**         Yellow Fever
56             0              4443**         Tuberculosis
32             1              4445**         Influenza
50-70          0              5110**         Lung Cancer
65             0              4110**         Diarrhea
45             1              4110**         Perinatal Complications
43             0              4445**         Whooping Cough
50-70          0              5110**         Lung Cancer
20-60          0              4445**         Cholera
```

Figure 5.2 Final anonymized data

5.3 Analysis

The existing method is improved at various phases of proposed approach. The analysis parameters are not quantified.

- **Privacy of the person:** However, it is clear from the discussion that the privacy of the person is protected. In fact it improved as there are positive changes made in its favor.
- **Data mining purpose:** The data mining purpose is very well full filled. More general view of attributes makes it easy for data mining techniques to use it.
- **Efficiency:** The efficiency of the algorithm is improved by making necessary changes and introducing a different ways.

CHAPTER 6

CONCLUSION AND FUTURE SCOPE

The problem of optimal anonymization method is NP-hard. The topic of privacy preserving data mining has much possibility to explore. The need for privacy preservation has studied which is emerged as latest research in data mining. The research work presented explains the various types of privacy preserving data mining techniques. The randomization process comes as the first one to be implemented for privacy preservation. The data perturbation method of randomization is mostly used one for data distortion. The method of k-anonymity includes processes of generalization and suppression to achieve anonymization. There are limitations to this method which are homogeneity attack and background knowledge attack. These shortcomings are overcome by the extension to it. L-diversity and t-closeness are the extension to the k-anonymity. Cryptographic techniques used various key based and also non-key based algorithms to secure data. These techniques add more overhead though. Soft computing model uses artificial intelligence techniques to for privacy preservation purposes. Lastly the distributed privacy preservation is reviewed for its contribution to privacy preservation.

The proposed work uses the combination of two techniques for privacy preservation. The randomization and k-anonymity are used in combine to transform data and secure it. The three important parameters data mining purpose, privacy, and efficiency are taken into consideration while implementing the two techniques. The dataset used in this thesis of medical institute or hospital. It is syntactically made. The modification are made to the existing process to enhance it and got the enhanced result too. As this topic is recent, there is much to do in various possible ways. In future other new techniques will be introduced which further enhance the security of sensitive data. Also, these techniques will be tested against synthetic as well as real time data.

There can be other combination of techniques too, such as randomization and cryptography, k-anonymity with soft computing, etc. More and more data sets which are having sensitive information will be processed through the improved method for privacy preserving data mining.

References

Abitha, N., Sarada, G., Manikandan, G., & Sairam, N. (2015, March). A cryptographic approach for achieving privacy in data mining. In *2015 International Conference on Circuit, Power and Computing Technologies (ICCPCT),* (pp. 1-5). IEEE.

Basu, A., Nakamura, T., Hidano, S., & Kiyomoto, S. (2015, August). k-anonymity: Risks and the Reality. In *Trustcom/BigDataSE/ISPA, 2015 IEEE*(Vol. 1, pp. 983-989). IEEE.

Bettini, S. M. C., Wang, X. S., & Jajodia, S. (2006). k-Anonymity in databases with timestamped data.

Blosser, G., & Zhan, J. (2007, August). Maintaining K-Anonymity on Real-Time Data. In *2007 International Conference on Machine Learning and Cybernetics* (Vol. 5, pp. 3012-3015). IEEE.

Burke, M. J., & Kayem, A. V. (2014, May). K-Anonymity for Privacy Preserving Crime Data Publishing in Resource Constrained Environments. In *2014 28th International Conference on Advanced Information Networking and Applications Workshops (WAINA),* (pp. 833-840). IEEE.

Chaum, D., Crépeau, C., & Damgard, I. (1988, January). Multiparty unconditionally secure protocols. In *Proceedings of the twentieth annual ACM symposium on Theory of computing* (pp. 11-19). ACM.

Chidambaram, S., & Srinivasagan, K. G. (2014, April). A combined random noise perturbation approach for multi-level privacy preservation in data mining. In *2014 International Conference on Recent Trends in Information Technology (ICRTIT),* (pp. 1-6). IEEE.

Clifton, C., Kantarcioglu, M., Vaidya, J., Lin, X., & Zhu, M. Y. (2002). Tools for privacy preserving distributed data mining. *ACM Sigkdd Explorations Newsletter, 4*(2), 28-34.

Dhanalakshmi, M., & Sankari, E. S. (2014, February). Privacy preserving data mining techniques-survey. In *2014 International Conference on Information Communication and Embedded Systems (ICICES),* (pp. 1-6). IEEE.

Du, W., & Atallah, M. J. (2001, September). Secure multi-party computation problems and their applications: a review and open problems. In *Proceedings of the 2001 workshop on new security paradigms* (pp. 13-22). ACM.

Elakkiya, R. T., & Velvizhi, P. (2013, February). Empowering privacy based multi-level trust using random perturbation techniques. In *2013 International Conference on Information Communication and Embedded Systems (ICICES),* (pp. 551-554). IEEE.

Gong, Z., Sun, G. Z., & Xie, X. (2010, May). Protecting privacy in location-based services using k-anonymity without cloaked region. In *2010 Eleventh International Conference on Mobile Data Management* (pp. 366-371). IEEE.

Han, J., Yu, H., & Yu, J. (2008, August). An improved l-diversity model for numerical sensitive attributes. In *2008. Third International Conference on Communications and Networking in China, 2008. ChinaCom* (pp. 938-943). IEEE.

Hellani, H., Kilany, R., & Sokhn, M. (2015, October). Towards internal privacy and flexible K-anonymity. In *2015 International Conference on Applied Research in Computer Science and Engineering (ICAR),* (pp. 1-2). IEEE.

Jian-min, H., Ting-ting, C., & Hui-qun, Y. (2008, May). An improved V-MDAV algorithm for l-diversity. In *2008 International Symposiums on Information Processing (ISIP),* (pp. 733-739). IEEE.

Kaur, R., & Bansal, M. (2015, September). Transformation approach for boolean attributes in privacy preserving data mining. In *2015 1st International Conference on Next Generation Computing Technologies (NGCT),* (pp. 644-648). IEEE.

Li, J., & Liu, G. (2011, December). On the Representation and Querying of Sets of Possible Worlds in the K-anonymity Privacy Protecting Model. In *2011 Seventh International Conference on Computational Intelligence and Security*.

Li, N., Li, T., & Venkatasubramanian, S. (2007, April). t-closeness: Privacy beyond k-anonymity and l-diversity. In *2007 IEEE 23rd International Conference on Data Engineering* (pp. 106-115). IEEE.

Li, X., Yan, Z., & Zhang, P. (2014, September). A review on privacy-preserving data mining. In *IEEE International Conference on Computer and Information Technology (CIT), 2014* (pp. 769-774). IEEE.

Liu, L., & Thuraisingham, B. (2006, December). The applicability of the perturbation model-based privacy preserving data mining for real-world data. In *Sixth IEEE International Conference on Data Mining-Workshops (ICDMW'06)* (pp. 507-512). IEEE.

Machanavajjhala, A., Kifer, D., Gehrke, J., & Venkitasubramaniam, M. (2007). l-diversity: Privacy beyond k-anonymity. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, *1*(1), 3.

Malik, M. B., Asger, M., Ali, R., & Sarvar, A. (2015, March). A model for privacy preserving in data mining using Soft Computing techniques. In *2015 2nd International Conference on Computing for Sustainable Global Development (INDIACom),* (pp. 181-186). IEEE.

Naor, M., & Pinkas, B. (2001, January). Efficient oblivious transfer protocols. In *Proceedings of the twelfth annual ACM-SIAM symposium on Discrete algorithms* (pp. 448-457). Society for Industrial and Applied Mathematics.

Pinkas, B. (2002). Cryptographic techniques for privacy-preserving data mining. *ACM Sigkdd Explorations Newsletter*, *4*(2), 12-19.

Samarati, P. (2001). Protecting respondents identities in microdata release.*IEEE transactions on Knowledge and Data Engineering*, *13*(6), 1010-1027.

Saranya, K., Premalatha, K., & Rajasekar, S. S. (2015, February). A survey on privacy preserving data mining. In *2015 2nd International Conference on Electronics and Communication Systems (ICECS),* (pp. 1740-1744). IEEE.

Sharma, M., Chaudhary, A., Mathuria, M., Chaudhary, S., & Kumar, S. (2014, July). An efficient approach for privacy preserving in data mining. In *2014 International Conference on Signal Propagation and Computer Technology (ICSPCT),* (pp. 244-249). IEEE.

Shen, Y., Liu, Y., & Zhang, Y. (2009, December). Personalized-Granular k-Anonymity. In *2009 International Conference on Information Engineering and Computer Science* (pp. 1-4). IEEE.

Sweeney, L. (2002, May). Achieving k-anonymity privacy protection using generalization and suppression. In *International Journal on Uncertainty, Fuzziness and Knowledge-based Systems*, 10 (5), 2002; 571- 588. IEEE

Tian, H., & Zhang, W. (2009, April). Extending l-diversity for better data anonymization. In *Sixth International Conference on Information Technology: New Generations, 2009. ITNG'09.* (pp. 461-466). IEEE.

Usha, P., Shriram, R., & Sathishkumar, S. (2014, February). Sensitive attribute based non-homogeneous anonymization for privacy preserving data mining. In *2014 International Conference on Information Communication and Embedded Systems (ICICES),* (pp. 1-5). IEEE.

Vijayarani, S., Tamilarasi, A., & Sampoorna, M. (2010, December). Analysis of privacy preserving k-anonymity methods and techniques. In *2010 International Conference on Communication and Computational Intelligence (INCOCCI),* (pp. 540-545). IEEE.

Wang, S. L., Tsai, Y. C., Kao, H. Y., & Hong, T. P. (2011, November). K-anonymity on sensitive transaction items. In *2011 IEEE International Conference on Granular Computing (GrC),* (pp. 723-727). IEEE.

Wu, Y., Sun, Z., & Wang, X. (2009, March). Privacy Preserving k-anonymity for Re-publication of Incremental Datasets. In *2009 WRI World Congress on Computer Science and Information Engineering,* (Vol. 4, pp. 53-60). IEEE.

Zacharouli, P., Gkoulalas-Divanis, A., & Verykios, V. S. (2007, April). A k-anonymity model for spatio-temporal data. In *2007 IEEE 23rd International Conference on Data Engineering Workshop,* (pp. 555-564). IEEE.

Zhang, X., & Bi, H. (2010, October). Research on privacy preserving classification data mining based on random perturbation. In *2010 International Conference on Information, Networking and Automation (ICINA)*(Vol. 1, pp. V1-173). IEEE.

Zhu, Y., & Peng, L. (2007, June). Study on k-anonymity models of sharing medical information. In *2007 International Conference on Service Systems and Service Management* (pp. 1-8). IEEE.