We are IntechOpen, the world's leading publisher of Open Access books Built by scientists, for scientists

4,800
Open access books available

122,000
International authors and editors

135M
Downloads

Our authors are among the

154
Countries delivered to

TOP 1%
most cited scientists

12.2%
Contributors from top 500 universities

CLARIVATE ANALYTICS BOOK CITATION INDEX INDEXED

WEB OF SCIENCE™

Selection of our books indexed in the Book Citation Index in Web of Science™ Core Collection (BKCI)

Interested in publishing with us?
Contact book.department@intechopen.com

Numbers displayed above are based on latest data collected.
For more information visit www.intechopen.com

Chapter 6

A Generic Scaffold Housing the Innovative Modus Operandi for Selection of the Superlative Anonymisation Technique for Optimized Privacy Preserving Data Mining

J. Indumathi

Additional information is available at the end of the chapter

http://dx.doi.org/10.5772/49982

1. Introduction

The genesis of novel data mining techniques has given an impetus to the privacy risks which has kept growing leaps and bounds. This is unquestionably plausible, owing to the probability fact to strappingly coalesce and interrogate gigantic data stores accessible on the web, in the midst of fumble of prior mysterious out of sight patterns. Privacy issues also augment its convolution because of new upcoming technologies; which are linking enormous number of reciprocally strange and erratic people, to make a worldwide economy. This scenario is of serious apprehension challenging consideration. This state of situation is like a life dealing with lemons. As learned folks we have to make lemonade, from the lemons necessitating this research, where the outcome is a generic scaffold.

The principal stipulation wished-for privacy is a high-quality data fortification modus operandi known as data mystification techniques. The importance of Privacy Preserving Data Mining (PPDM) down to the core is hackneyed not only from its ostentation to heave out crucial knowledge, but also from its resistance to assault*PPDM is a discipline whose desire is to authorize delivery transmits of respondent data while preserving respondent privacy.*It introduces solutions to problems where the question is how to get hold of data mining results without violating privacy.

In the proposition of novel solutions there are two vital things to be highlighted. The first one is *Privacy* of users and personal data within strenuous environments and implicit communities. The second one is *Information Security* as it relates to privacy and the information resources provided in the same environments. This chapter aspires to

contribute to the solution of a specific problem, namely, the problem of sharing sensitive data. Developing new, improvising existing algorithms and techniques for PPDM is endeavored.

1.1. Motivation

In the midst of the implausible detonation of data garnering, data dissemination, internet technologies and the manifestation of susceptible applications, security issues in knowledge discovery have reached the pinnacle. The chic data analysis and mining techniques have produced a sharp awareness of their potential, compromising privacy, and has posed challenges of controlling data access in accordance with privacy policies. The escalating exploitation done by them have reached the zenith, and have surfaced as a central and ubiquitous problem. To safeguard knowledge and wisdom, which, is pre-eminence as can be seen in figure 1, but as humble users of the most modern technologies we are pitted with possessions that may even make us paranoidconcerning usage of a computer. This chapter will bring to the surface, the modus operandi needed to shield privacy.

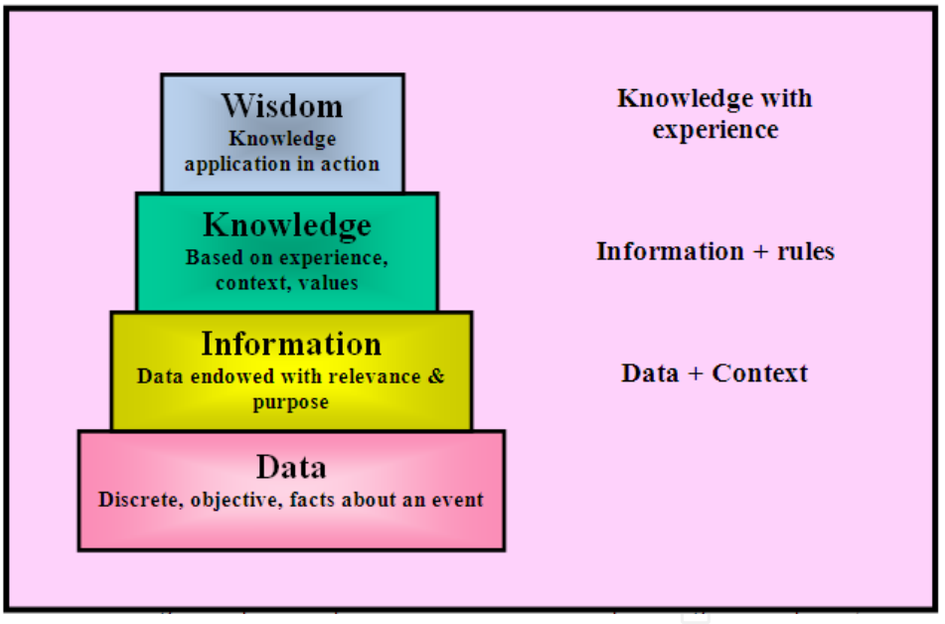

Figure 1. Data Pyramid

The objective of Privacy-Preserving Data Mining(PPDM)as can be seen in figure 2, is to release a privacy preserved dataset which will spot and protect the sensitive information in the data (with high probability),so that researchers can study the data without compromising privacy of any individual. The task of PPDM force's a central thrust on establishing a world with robust data security, where knowledge users persist to profit from data without compromising the data privacy (Indumathi, J. et al.,(2007a))

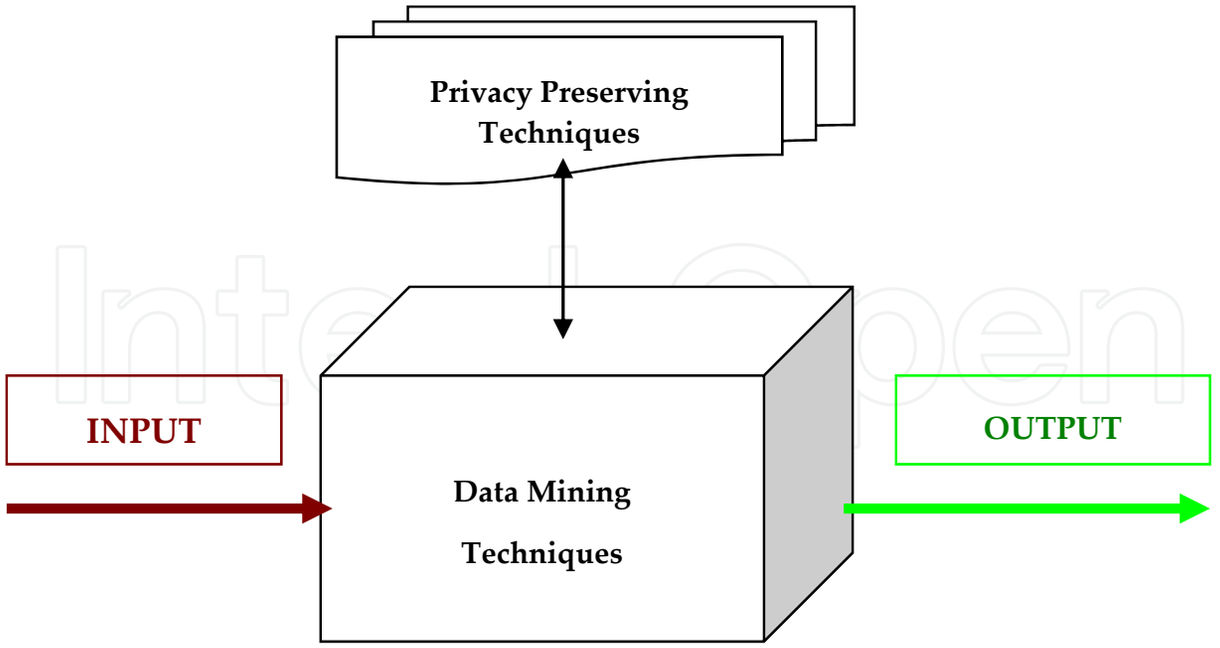

Figure 2. A Framework for Privacy Preserving Data Mining Systems: High-Level

PPDM like data mining is an interdisciplinary field, as can be seen in figure 1.3,involving the confluence of a set of disciplines, including database technology, statistics, machine learning, visualization, and information science etc., Not only that it is also divided based on the data mining approaches used, the kinds of data to be mined or on the given data mining application, the data mining system may also integrate techniques from spatial data analysis, image analysis, signal processing, computer graphics, economics, business, bioinformatics, or psychology etc.,

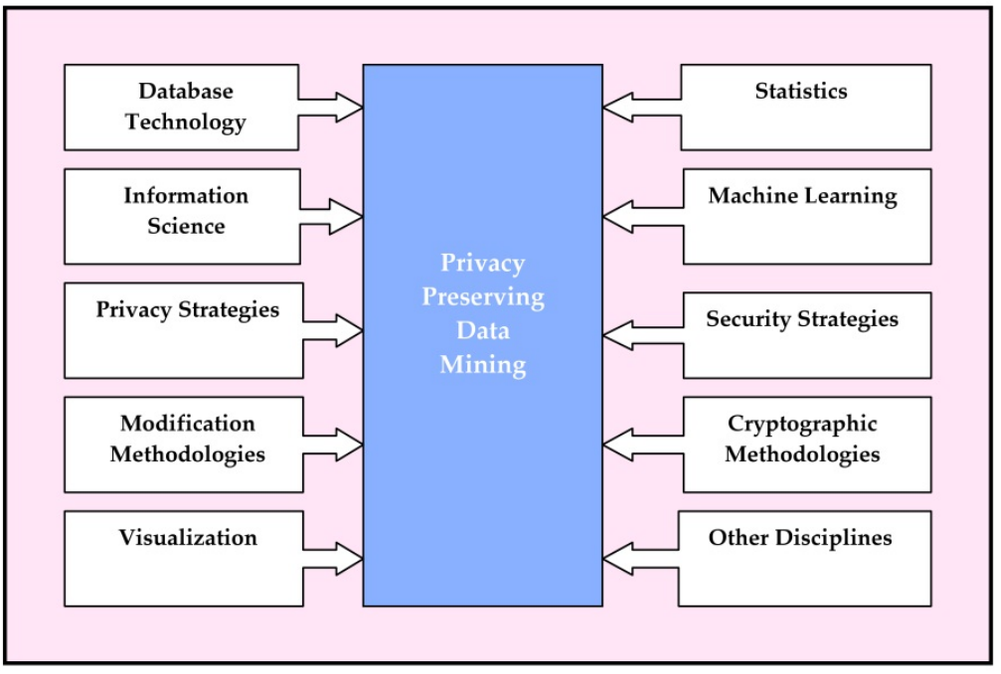

Figure 3. PPDM Confluences

2. Literature survey

This part deals with a meticulous survey bringing to limelight scores of works on the various existing privacy preserving techniques, their advantages and deficiencies. The majority techniques for privacy computation use some form of transformation on the data in order to perform the privacy preservation. Characteristically, such methods decrease the granularity of representation or limit access to resources in order to reduce the privacy. This dwindling in granularity results in some trouncing of efficacy of data mining algorithms. This is the normal trade-off between information loss and privacy. Researchers developed methods to enable data mining techniques to be applied while preserving the privacy of individuals. Though several approaches have been proposed for privacy preserving data mining, at this point we would like the reader to read Verykios et al (2004 b), Mohammad Reza Keyvanpour et al.,(2011) for a quick overview. Verykios et al (2004 b) gives a detailed survey on some of the techniques used for PPDM. Mohammad Reza Keyvanpour et al.,(2011) proposed a classification based on three common approaches of Privacy Preserving data mining, namely, Data modification approach, Data sanitization approach and Secure Multi-party Computation approach.

Atallah et al. (1999)considered the problem of limiting disclosure of sensitive rules[Statistical Disclosure Control(SDC)], aiming at selectively hiding some frequent itemsets from large databases with as little impact on other, nonsensitive frequent itemsets as possible. Specifically, the authors dealt with the problem of modifying a given database so that the support of a given set of sensitive rules, mined from the database, decreases below the minimum support value.

Most data mining techniques, i.e. association rule mining and classification, are well studied by followers of both approaches. Agarwal et al.,(2000) and Vassilios et al.,(2004) for data sanitization techniques; Kantarcioglu et al.,(2004), Vaidya et al.,(2002) and Du, W., Zhan (2002) are based on secure multi-party computation techniques.

Jian Yin et al.,(2005), proposed model-based solutions for the privacy preserving clustering problem. Data holder parties build local models of their data that is subject to privacy constraints. Then a third party builds a global model from these local models and cluster the data generated by this global model. All of these works follow the sanitization approach and therefore trade-off accuracy versus privacy. Except Jian Yin et al.(2005), none of them address privacy preserving clustering on horizontally partitioned data. Privacy preserving techniques for clustering over vertically partitioned data was proposed by Vaidya et al.(2002) . Ali Inan et al.,(2006) and Oliveira et al.,(2004) mainly concentrate on finding object based dissimilarity for privacy preservation.

Having thoroughly investigated the available diverse techniques for privacy preservation, we find that *the level of Privacy Preservation involved is* only *of* a single level. *Even the newly proposed Privacy Preservation techniques*([i.e., A comparative study on the various cryptographic methods for future work(Vasudevan.V et al. 2007); the flustering *technique*(Indumathi.J et al.2007c)] was implemented and evaluated on our own conceptual

frameworks(Indumathi.J et al. 2007a,b, Gitanjali.J et al. 2007, Indumathi.J et al. 2008a,b*) to prove their efficiency. They are also of single level ones. The framework* was used to compare and contrast each and every one of the techniques in a general podium which will be the basis for ascertaining the suitable technique for a given type of application of privacy preserving shared filtering. Nonetheless, there are situations where the sharing of data can lead to general gain as in the case of privacy preserving secure accord as mentioned (Indumathi.J et al.,2007b*).*

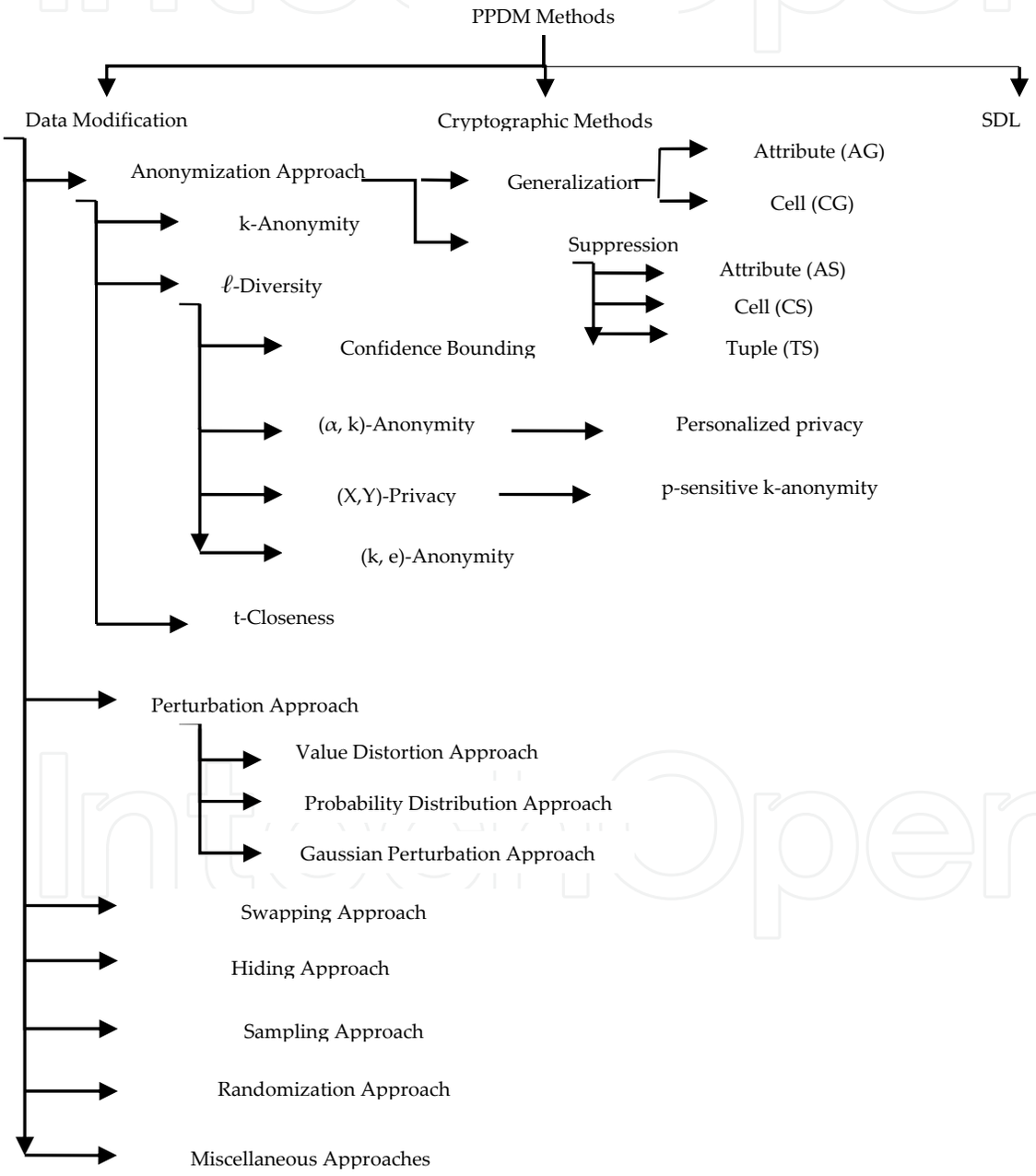

Figure 4. Proposed PPDM Taxonomy

3. Proposed taxonomy

In spite of the different categorizations available for PPDM, we propose a new taxonomy based on the procedures involved as shown in figure 4. Our earlier proposed taxonomy for PPDM (Indumathi.J et al. 2007c) is further subdivided into perturbation-based, swapping-based, hiding -based, sampling -based, randomization -based, anonymization-based techniques etc.,

PPDM Methods are broadly categorized into Data Modification based methods, cryptographic based methods; Statistical Disclosure Limitation (SDL) based methods. The data modification methods are further subdivided into perturbation-based, swapping-based, hiding -based, sampling -based, randomization -based, anonymization-based techniques etc., (Indumathi.J et al. 2007c). We will pin down onto the anonymization-based techniques.

3.1. Significance of study for anonymisation based technique

Anonymity is a derived from the Greek word α ′νωνυμ΄ια, *anonymia*, which means "without a name" or "namelessness". In conversational use, anonymous typically refers to a person, and often means that the personal identity, or personal identifiable information of the given person is unidentified. Anonymityis frequently observed as the *superlative way to protect individual privacy in the biomedical background*. For instance, numerous cases in biomedical law prove that anonymity can serve as a warranty for self-sacrifice. It is also seen as a protection against scientific and other biases in research. Contributing an organ donation/blood to an identified person and to receive an organ donation/blood from an identified person creates a special and complicated interpersonal relationship between the donor and the recipient. Amongst the various techniques the identities of the donor and recipient are hidden through anonymisation.

3.2. Sub-classification of anonymisation based technique

Anonymization is a process that confiscates identity information from a communication or record by making it pseudonymous, in which case the same subject will always have the same replacement identity but cannot be identified as an individual.

3.2.1. Generalization & suppression

3.2.1.1. Generalization

Generalization consists of replacing with attribute values with semantically consistent but less precise values. For example, the place of birth can be replaced by the country of birth which occurs in more records so that the identification of a specific individual is more difficult. Generalization maintains the correctness of the data at the record level but results in less specific information that may affect the accuracy of machine learning algorithms applied on the k-anonymous data set. Different systems use various methods for selecting the attributes and records for generalization as well as the generalization technique [Friedman.A,et al.,(2008)].

Generalization can be applied at the following levels:

- **Attribute (A_G):** In this sub-type, generalization is executed at the level of column; a generalization step generalizes all the values in the column.
- **Cell (C_G):** In this sub-type, generalization is executed on single cells; as a result a generalized table may contain, for a specific column, values at different generalization levels. For instance, in the date of birth column.

3.2.1.2. Suppression

Suppression refers to removing a certain attribute value and replacing occurrences of the value with a special Value?, indicating that any value can be placed instead. Suppression can drastically reduce the quality of the data if not properly used [S.V. Iyengar(2002)].

Suppression can be applied at the following levels

- **Tuple (T_S):** In this sub-type, suppression is executed at the level of row; a suppression operation removes a whole tuple.
- **Attribute (A_S):** In this sub-type, suppression is executed at the level of column; a suppression operation obscures all the values of the column.
- **Cell (C_S):** In this sub-type, suppression is executed at the level of single cells; as a result a k-anonymized table may wipe out only certain cells of a given tuple/attribute.

3.2.2. K-anonymity technique

In this technique each record within an anonymized table must be indistinguishable with at least k-1 other record within the dataset, with respect to a set of QI attributes or if one record in the table has some QID, at least k-1 other record also have the value QID. QID should have at least k minimum group size value. In particular, a table is K-anonymous if the QI attributes values of each record are identical to those of at least k-1 other records. To achieve the K-anonymity requirement, generalization or suppression could be used [P. Samarati et al.,(1998), L. Sweeney(2002)].

Advantages

- Individual record hidden in a crowd of size k

Disadvantages

- Identifying a proper QID is a hard problem.
- Finding a k-anonymity solution with suppressing fewest cells

Attacks

- Homogeneity and background knowledge attack.
- Record linkage.

3.2.3. ℓ-diversity

An equivalence class is said to have ℓ-diversity if there are at least one "well-represented" values for the sensitive attribute. i.e., ℓ-diversity requires every QID group to contain at

least one "well represented" sensitive values.ℓ-Diversity provides privacy preserving even when the data publisher does not know what kind of knowledge is possessed by the adversary.

The main idea of ℓ-diversity is the requirement that the values of the sensitive attributes are well-represented in each group. The *k*-anonymity algorithms can be adapted to compute ℓ-diverse tables [A.Machanavajjhala et al,(2006)]. ℓ-Diversity resolved the shortcoming of *k*-anonymity model.

Advantages

- It prevents homogeneity and background knowledge attack.

Disadvantages

- ℓ-diversity may be difficult and unnecessary to achieve.
- ℓ-diversity is insufficient to prevent attribute disclosure (i.e., it tends to skewness and similarity attack).
- *Distinct* ℓ-diversity does not prevent probabilistic attack.

3.2.3.1. Confidence bounding

Wang et al. [2005, 2007] considered bounding the confidence of inferring a sensitive value from a *QID* group by specifying one or more *privacy templates* of the form, $QID \rightarrow s, h$; *s* is a sensitive value, *QID* is a quasi-identifier, and *h* is a threshold.

Advantages

- It allows the flexibility for the data publisher to specify a different threshold *h* for each combination of *QID* and *s* according to the perceived sensitivity of inferring *s* from a group on *QID*

Disadvantages

- It does not prevent attribute linkage attack.

3.2.3.2. (α, k)-Anonymity

Wong et al. [2006] proposed a similar integrated privacy model, called (α, *k*)-*anonymity*, requiring every *QID* in a Table *T* to be shared by at least *k* records and $conf(QID \rightarrow s) \leq \alpha$ for any sensitive value *s*, where *k* and α are data publisher-specified thresholds. Nonetheless, both (*X,Y*)-Privacy and (α, *k*)-anonymity may result in high distortion if the sensitive values are skewed.

3.2.3.3. (X,Y)-Privacy

(*X,Y*)-anonymity states that each group on *X* has at least *k* distinct values on *Y* (e.g., diseases). However, if some *Y* values occur more frequently than others, the probability of inferring a particular *Y* value can be higher than 1/*k*. To address this issue, Wang and Fung [2006] proposed a general privacy model, called (*X,Y*)-Privacy, which combines both (*X,Y*)-

anonymity and confidence bounding. The general idea is to require each group x on X to contain at least k records and $conf\ (x \rightarrow y) \leq h$ for any $y \in Y$, where$Y$ is a set of selected sensitive values and h is a maximum confidence threshold.

3.2.3.4. (k, e)-Anonymity

Most work on k-anonymity and its extensions assumes categorical sensitive attributes. Zhang et al. [2007] proposed the notion of (k, e)-anonymity to address numerical sensitive attributes such as salary. The general idea is to partition the records into groups so that each group contains at least k different sensitive values with a range of at least e. However, (k, e)-anonymity ignores the distribution of sensitive values within the range λ.

3.2.4. t-Closeness

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. ℓ-diversity principle represents an important step beyond k-anonymity in protecting against attribute disclosure. However, it has several shortcomings, such as ℓ-diversity may be difficult and unnecessary to achieve and it is insufficient to prevent attribute disclosure. N. Li et al.,(2007) proposes a new privacy measure is called t-Closeness, which requires that the distribution of a sensitive attribute in any equivalence class is close to the distribution of the attribute in the overall table.

Advantages

- It prevents skewness attack.
- It provides efficient closeness by using earth mover distance.

Disadvantages

- It lacks in flexibility specifying different protect levels for different sensitive value.
- EMD function is not suitable for preventing attribute linkage on numerical sensitive attribute
- It greatly degrades the data utility because it requires the distribution of sensitive attributes to be the same in all QID groups.

Diving down into the contemporary literature with the intention of hauling pearl, one finds that only a handful faction have navigated the area but no one has ever projected anapt explication or endeavored their implementation in the real world..Having discussed all the existing literature on the Optimality Criteria, definitions of safety (privacy), validity (utility) we can now try to illustrate how they are used for the comparison of alternative data releases and how we can customize our Modus Operandi for selection of the best PPDM Technique for any real–time Application. Moreover, we have already implemented and evaluated the true efficiency of the PPDM techniques [(Indumathi.J et al.,2007b,2008d,2009)] on our own conceptual frameworks [Indumathi.J et al.,2007,2008a].In this paper we have attempted to structure and suggest the Modus Operandi for Selection of The Best Anonymity based PPDM Technique for any real time application.

4. Problem statement

Prescribe a new generic scaffold which will obscure data sets (while still preserving certain statistical characteristics of the data) by choosing the best **anonymisation** technique with the intention of trying to improve the level of privacy protection and to maintain a perfect balance between privacy and utility.

4.1. Problem description

We have determined (Indumathi.J et al., (2009)to put forward a new generic scaffold solution for Data-Garnering and PPDM System as shown in figure 5.and incorporate five aims, in five steps namely:

a. An improved preference mechanism to determine the best optional privacy-preserving method as shown in figure 5 not only for data mining but also in shielding the information resources;
b. Develop notation for assessment and evaluation of alternative concealed data PPDM;
c. Guard data in database structures and erstwhile forms of information possessions;
d. Apply the guideline, database principle to enforceable security policies and procedures; and Implement the solution and validate it in real world situation.

It is already (Indumathi.J et al., (2009)stated that the state of being free from unsanctioned intrusion is known privacy. It limits the risk of disclosure of confidential information. Utility of concealed data is to measure how closethe concealed data are to their unconcealed version. There is no general accord that assessment and evaluation of alternative concealed data ought to be done on the foundation of the privacyand utilitythat these produce; these tend to be ambiguous concepts.

5. Architecture of the proposed system

A privacy-enhancing three tier architecture that enables to preserve the privacy of data by combining three best methods for Privacy Preservation, namely access control limitation technique, randomization and Privacy Preserving Clustering (PPC) shows an increase in the performance by modifying the algorithm so that the dissimilarity matrix is found only once by the third party instead of it being found by each local party (Indumathi.J et al.,2007c).].

In this paper we have improvised on the access control mechanism by surrogating it with the Purpose Based Access Control (PBAC) which confers us with three benefits viz., First, PBAC models include features to establish role hierarchies; second, RBAC policies change very little over time; third, it naturally supports delegation of access permissions.

As depicted in Figure 5, our framework encompasses a transactional database (modelled into a text database), a set of algorithms used for flustering data from the database, a transaction retrieval engine for fast retrieval. We amend (Indumathi.J et al., (2009) and bring out a diagrammatic representation of the architecture as shown in figure 5 and 6 involved in the proposed architecture.

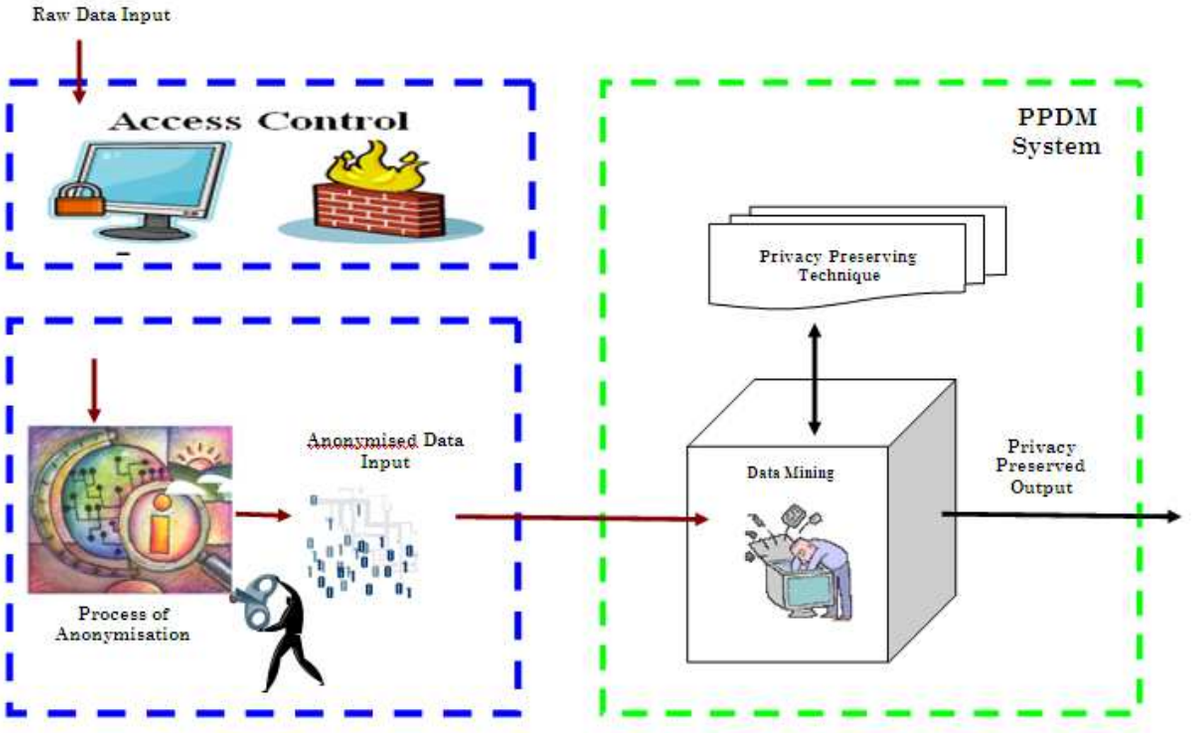

Figure 5. The PPDM Framework: High-Level

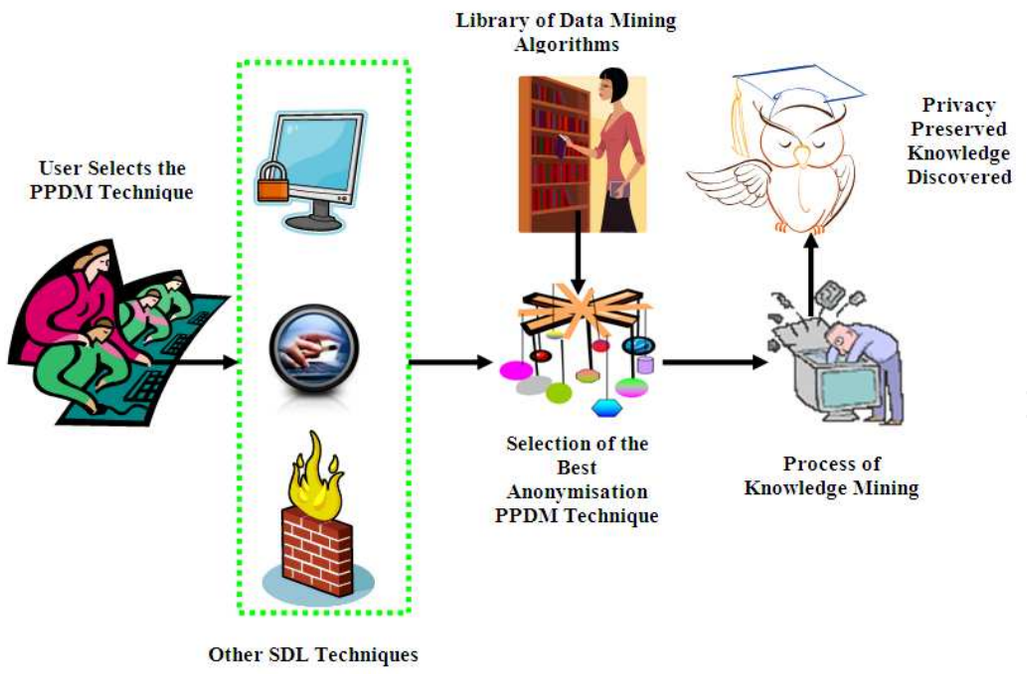

Figure 6. Modus Operandi for selecting the Best Anonymisation PPDM Technique for an Application-High Level Diagram.

Data concealingis used to limit the revelation of data by applying a mystification process to the originally collected data and publishing it the public (Indumathi.J et al., (2009). We refer synonymously the Mystification process of the original data as data befuddling; the terms

concealed data set represents the output of such mystification; the concealed data set refers both to the published data and the information provided to the users about the mystification.

Data collectors collect data from data providers. These heterogeneous data can be protected by limiting the data access using password, firewalls etc., Owing to the versatility of thedata mining tasks, any one best suitable technique can be selected from a family of privacy-preserving data mining (PPDM) methods using the technique selector. This technique is used for protecting privacy before data are shared. The technique selector can either use the PPDM Ontology to select the desired technique. The algorithms are used for modifying the unique facts by some means, with the intention that the private data and private knowledge linger private even subsequent to the mining process. Metrics is used to find the measure of these techniques. The input to this block is unpreserved data whereas its output is privacy preserved and befuddled data. This is given as an input and is subject to any of the data mining techniques.The figure 7 shows the Modus Operandi for selecting the Best PPDM Technique for an Application.

From a privacy-preserving data mining (PPDM) methods pool we select(Indumathi.J et al., (2009) any one best suitable anonymisation technique can be selected using the technique selector. This technique is used for protecting privacy before data are shared. The technique selector can either use the PPDM Ontology (Indumathi.J et al., (2009) to select the desired technique. The algorithms are used for modifying the unique facts by some means, with the intention that the private data and private knowledge linger private even subsequent to the mining process. Metrics is used to find the measure of these techniques. The input to this block is unpreserved data whereas its output is privacy preserved and befuddled data. This is given as an input and is subject to any of the data mining techniques.The figure 4.3 shows the Modus Operandi for selecting the Best anonymisation PPDM Technique for an Application.

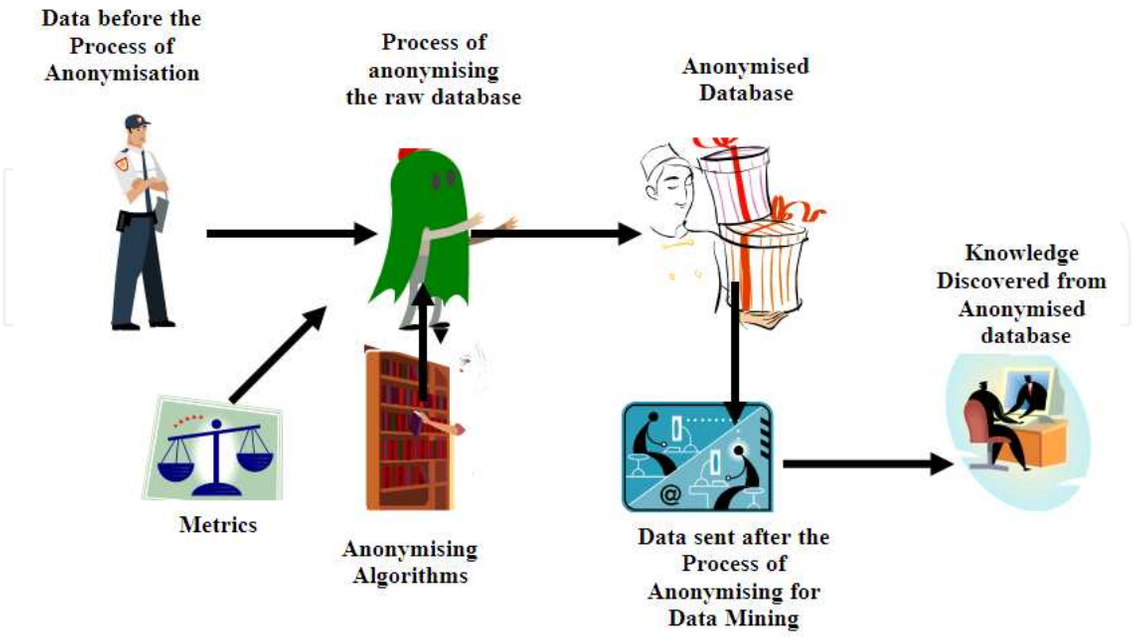

Figure 7. Modus Operandi for Selecting the Best Anonymisation PPDM Technique for an Application-Middle Level Diagram.

6. System architecture design

Problem Formulation: Specification of an appraisal framework in order to compare and contrast each and every one of the techniques in a general podium which will be the basis for ascertaining the suitable technique for a given type of application.

6.1. Subsystem architecture of the purpose based access control framework for PPDM systems

One main motive of the project work is to discuss how to address privacy issues, that is, how a business providing organization can ensure that the privacy of its consumers is protected. The project implements an approach based on intended purpose and access purpose. Intended purpose implies the intended usage of data. The purpose for which a data element is accessed is implied by access purpose. The purpose compliance check introduces some overhead. Two different granularity levels are used for purpose compliance and the overhead incurred in each scheme is noted. In addition, the approach is implemented on data that is reliable and of good quality in order to make important decisions based on that data. The concept of obtaining quality data using confidence policies is implemented in the project.

First, the emphasis is laid on the importance of privacy protection with respect to any business providing organization. While enforcing good privacy practice helps attract customers, there is a possibility of potential lawsuits by customers if the privacy of their data is violated. This prompts organizations to ensure and focus on privacy protection. One model that aims at protecting the privacy is purpose based access control for relational databases. In this model, purpose information is associated with a given data element. Such a privacy protecting access control model is desired since privacy protection cannot be easily achieved by traditional access control models .The concept of purpose is thus introduced in access control models. The proposed approach introduces access purpose and intended purpose. Intended purpose specifies the intended usage of data, and access purposes specifies the purposes for which a given data element is accessed. If privacy officers desire that data should not be allowed for certain purposes, this model can be used since it supports prohibitions. The granularity of data labeling, i.e., the units of data with which purposes can be associated is introduced using two schemes. In one scheme, purpose is associated with every data element in a relation and in the other; it is associated with each attribute in a relation. The role based access control (RBAC) is chosen as the underlying model to implement the purpose concept. RBAC, by itself, is a way of enforcing access to system resources for authorized users. Such a mechanism will be helpful in designing applications which aim at preserving data privacy.

6.1.1. High level design of purpose based access control framework for PPDM systems

The features of the proposed Purpose Based Access Control are as follows:

1. An access control model for privacy protection based on the notion of purpose was implemented.

2. The overhead incurred by using the purpose compliance checks was calculated for two labeling schemes using synthetic datasets.
   - The element based scheme (finest granularity) had significant overhead compared to the attribute based labeling.
   - Thus, the response time increased as granularity of labeling became finer.
3. Providing high quality data to decision makers using confidence values associated with data was implemented. A greedy algorithm was implemented which dynamically incremented data confidence values in order to return query results that satisfied the stated confidence policies.

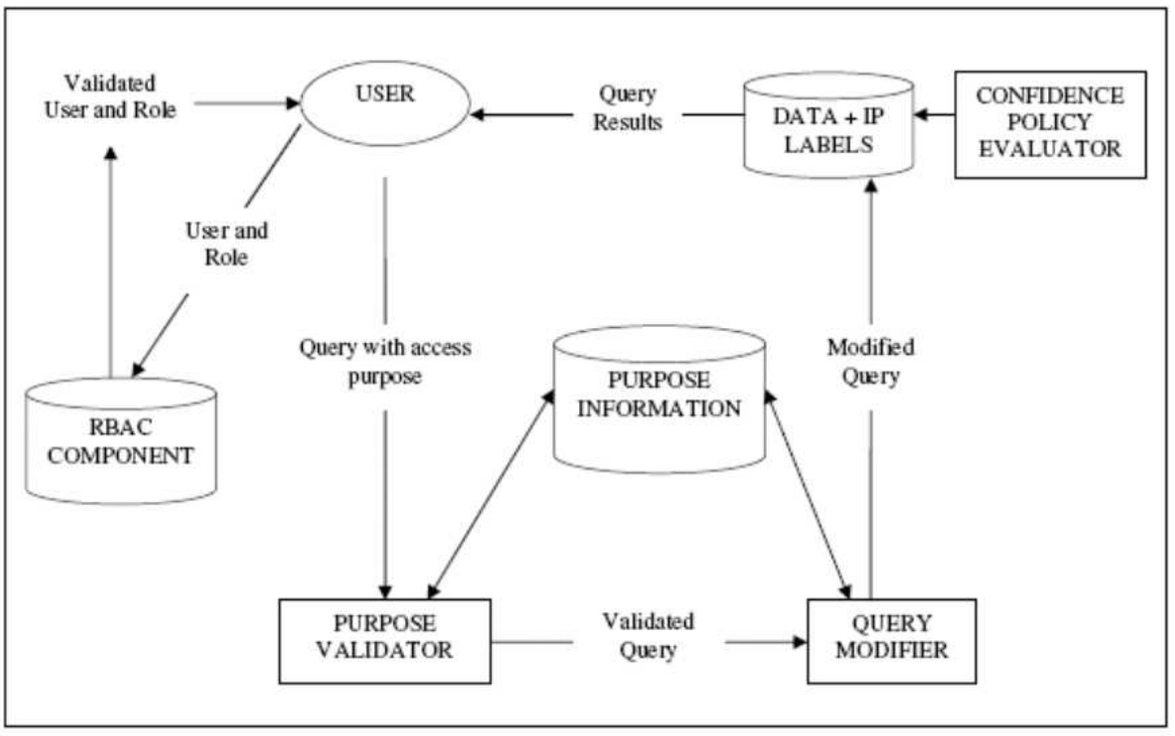

Figure 8. Block Diagram of the proposed Purpose Based Access Control Privacy Protection System

The details of the architecture are discussed. In figure 8, the overall picture of the privacy protection system is shown. The system's architecture consists of the six main components: The RBAC component, Purpose Management component, Query Modification mechanism, Confidence Policy Evaluation, Data which includes the intended purpose labels.

The RBAC component includes role and user creation, role assignment to users and role activation.

The Purpose Management component includes purpose storage and validation. In the purpose creation, we identify a purpose tree based on the organizational needs and create a purpose table to store the hexadecimal encodings of each purpose node. In purpose validation, we first find the conditional role of a user giving the role as an input. We also verify the access purpose of the user. The allowed intended purpose (AIP) and prohibited intended purpose (PIP) of a data element is also found. Finally, we check the compliance

between the access purpose given by the user and the intended purpose (AIP and PIP) of the data element as decided by the privacy policy.

The query modification component returns the results of the query after the purpose compliance check. If the access purpose is non-compliant with the intended purpose of the data element, the corresponding result of the data element should not be displayed.

The confidence policy evaluation ensures that the data is of good quality and reliable in order to make important decisions. It includes confidence assignment, whereby each data item in the concerned relations is assigned a confidence value of around 0.1, confidence evaluation of the query results, policy evaluation which returns only those results with confidence value higher than the threshold specified in the confidence policy to the user (one who is managing the data). Finally, strategy finding and data quality improvement can be used to dynamically increment the confidence level of data to return query results that satisfy the stated confidence policy.

The flow is as follows. The Confidence Policy Evaluation first improves the quality of the data which will later be used for the privacy protection process. Before the user can issue a query, he should be assigned user name and a role in the database. The query is then passed onto the purpose validator component.

6.2. Subsystem architecture of the framework for PPDM systems

The architecture of the proposed system has two major subsystems called Privacy Preserving Framework, Data Mining Subsystems.Our already implementedAgent Based PPDMT Selector Module of PPDM Methods (Indumathi.J et al., (2009) intelligently decides the best suited technique as it is the routine that waits in the background and performs an action when a specified event occurs.

The PPDM Framework sub system is designed as a generalized approach to support privacy-preserving data doling. Most work in this field has been extremely specialized by edifying a new all-purpose framework, we anticipate facilitating the application of privacy-preserving mechanisms and data doling to a broader variety of fields. Furthermore, this approach has been purposeful en route for the development of a practical combined infringement discovery structure. The methods in this thesis have been distinctively spear-headed for real-world solutions that establish privacy requirements.

6.2.1. Privacy preserving framework subsystem

The Privacy Preserving Framework subsystem [Indumathi.J et al.,(2007)]has components, namely, metrics module, Library of algorithms module, technique selector module. The Privacy Preserving Framework categorically decides the Privacy preservation mode of selection (manual or automatic or interactive) of technique from the technique selector module, with the help of Library of algorithms module. The metrics module is used to quantify the work.

Library of algorithms - presents the taxonomy of the anonymisation based PPDM algorithms.

Metrics Module - Evaluator sub system has four components, namely, the performance measurer, data utility measurer, level of uncertainty measurer, resistance measurer. A introductory list of appraisal parameters to be used for assessing the worth of privacy preserving data mining algorithms, is given below:

i. Performance of the proposed algorithms

The performance of the proposed algorithms in stipulations of time requirements, that is the time essential by each algorithm to conceal a particular set of susceptible information; assess the time needs in terms of the standard number of operations, essential to decrease the incidence of facade of explicit sensitive information below a specified threshold.

The *communication cost* incurred through the barter of information among a number of collaborating sites, should be painstaking. It is crucial that this cost must be reserved to a minimum for a distributed privacy preserving data mining algorithm.

ii. Data utility

*T*he *data utility* after the submission of the privacy preserving technique, which is equal with the minimization of the information loss or else the loss in the functionality of the data; the measure used to evaluate the information loss depends on the specific data mining technique with respect to which a privacy algorithm is performed. Information loss in the background of association rule mining, classification, will be calculated either in terms of the amount of rules that were both residual 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 clustering, the variance of the distances among the clustered items in the original database and the befuddled database, can be the basis for evaluating information loss in this case.

iii. Uncertainty level

The *level of uncertainty is defined as* with which the sensitive information that have been hidden can still be predicted; these privacy preservation strategies, demote the information to facilitate protection below certain thresholds.

iv. Fortitude of confrontation to diverse Data Mining techniques

The fortitude of confrontationisaccomplished by the privacy algorithms; to different data mining techniques. The aim of algorithms is the fortification of susceptible information against illegal disclosure. The intruders and data terrorists will try to conciliate information by using various data mining algorithms.

6.3. Subsystem architecture of the framework for data mining systems

The Data Mining subsystem as shown in figure 9. have five **c**omponents, namely, Database Server Module, Data Warehouse Server Module, Visualization Module, Data Mining

Module and Data Preprocessing Module. Data Mining Module has been confined only to three components, namely, association, classification, clustering.

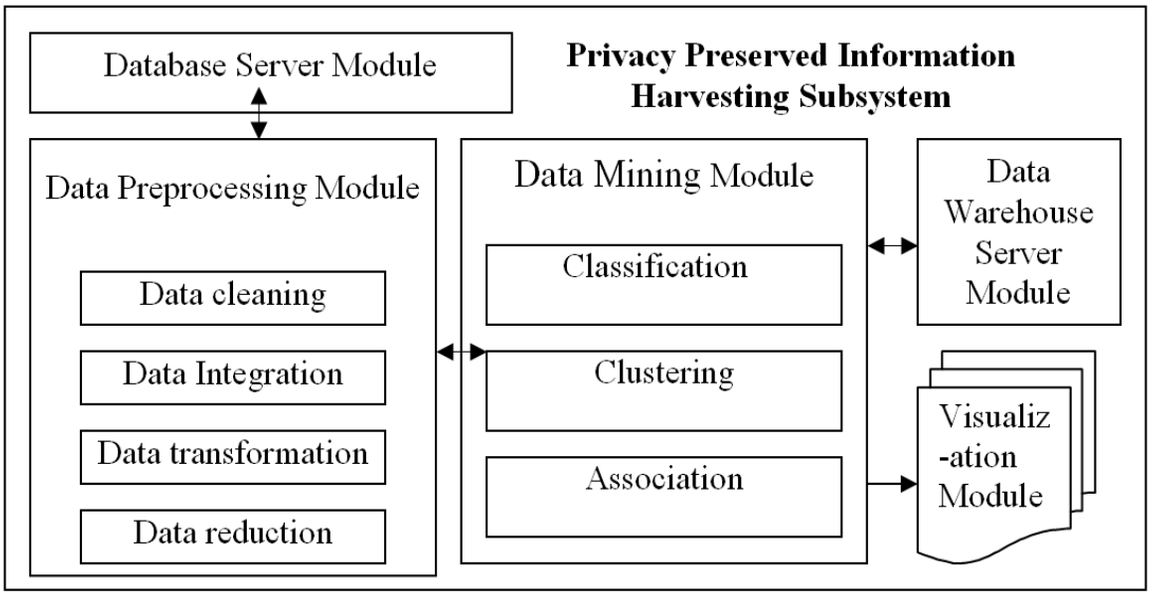

Figure 9. Privacy Preserving Data Mining Framework Subsystem.

7. Evaluation and experimental results

7.1. Data structure design and datasets used

The framework is tested and implemented for patient data collection in health care systems. The collected patient details are mined for finding the relationship between the eating habits and the diseases that are caused due to it.We have used the geneId dataset, and datasets collected from health care centres around Vellore and transmitted to Vellore City hospitals where the specialists advice the patients based on the outcomes.

User characteristics-The target user is expected to have knowledge about the data used in the system. The user should have experience in working with the standard windows environment.

Operating Environment-The operating environment of the software is in the Data Mining area at Health Care domain.

7.2. Metrics

Data Utility is the percentage of similarity between the data mined results from original data and randomized data.

Data Privacy: For quantifying privacy provided by a method, we use a measure based on how closely the original values of a modified attribute can be estimated. If it can be estimated with c % confidence that a value x lies in the interval [x1; x2], then the interval

width (x2 - x1) defines the amount of privacy at c % confidence level. Based on the above factor we are going to analyze the various anonymity approaches that have been implemented.

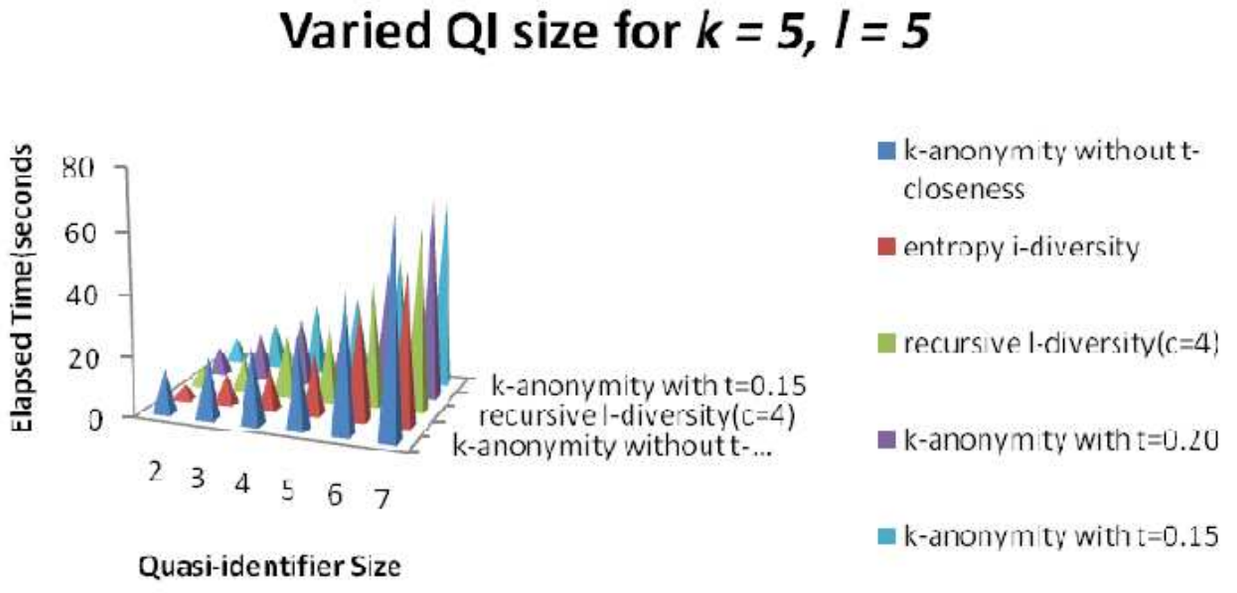

Figure 10. Elapsed time Vs. quasi-identifier size for diverse anonymity techniques

Figure 10. shows that as the data size /dimensions are increasing there is an decrease in the data utility.

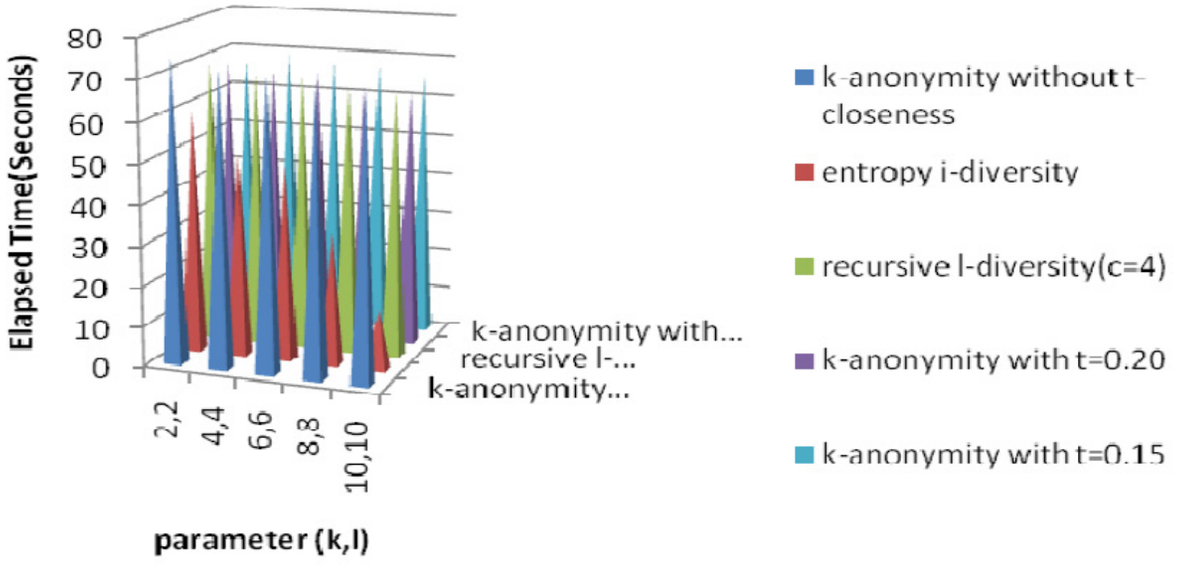

Figure 11. Elapsed time Vs. parameter for diverse anonymity techniques

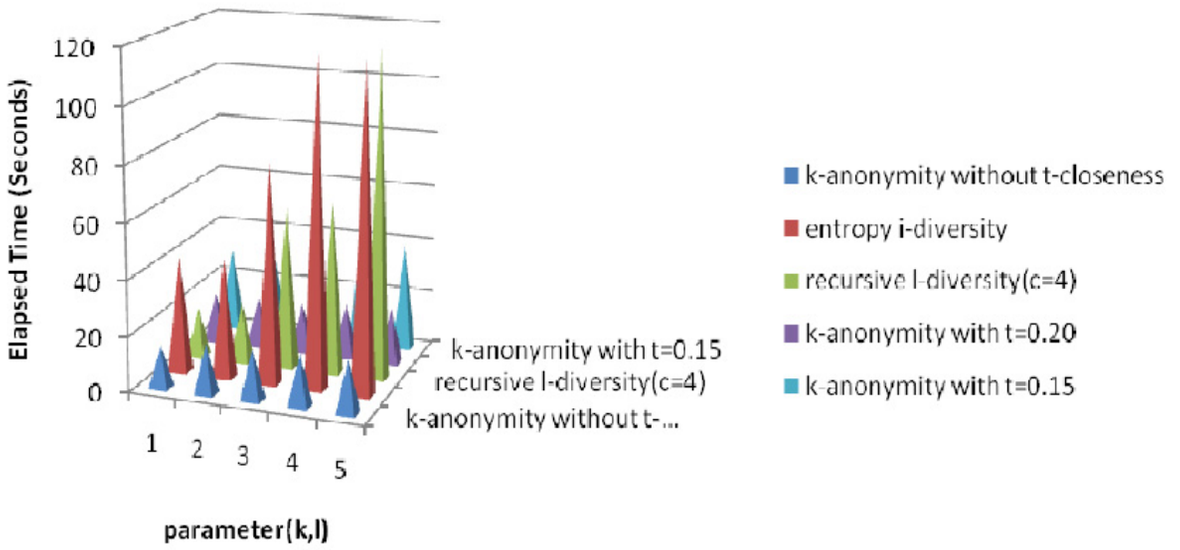

Figure 12. Elapsed time Vs. parameter for diverse anonymity techniques

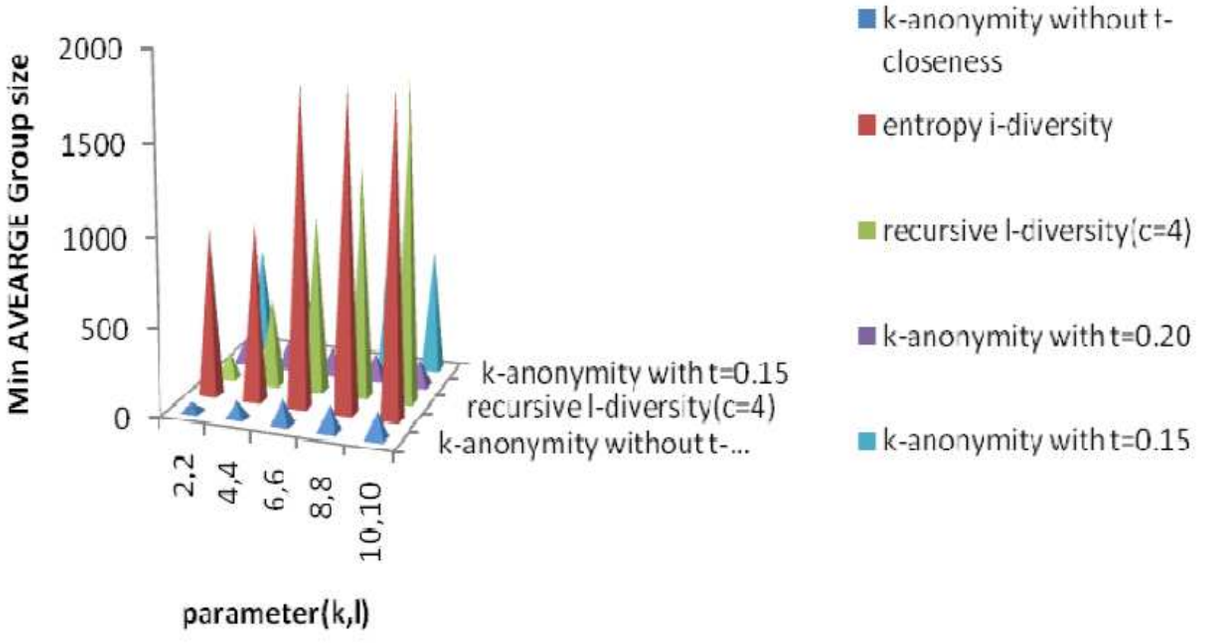

Figure 13. Elapsed time Vs. parameter for diverse anonymity techniques

Figure 11, 12, 13. shows Elapsed time Versus parameter for diverse anonymity techniques.

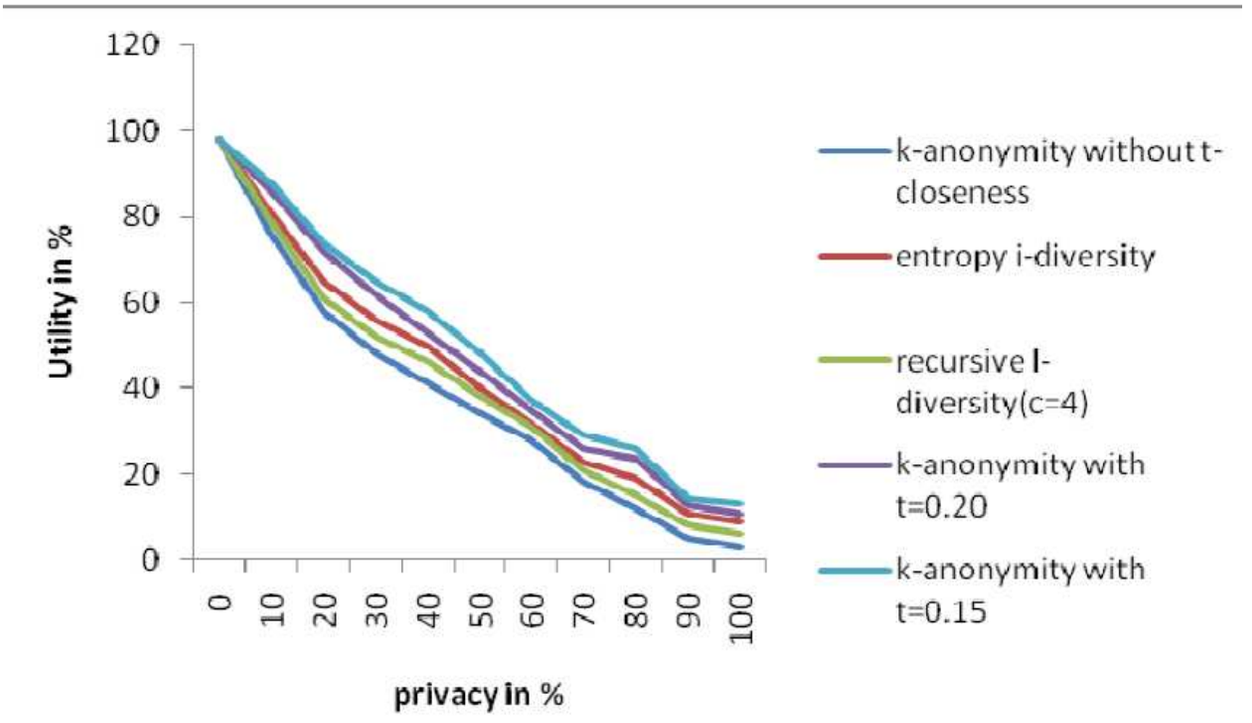

Figure 14. Data Utility Versus Privacy for diverse anonymity techniques

Figure 14. shows that as the **Privacy** increases there is an decrease in the data utility.

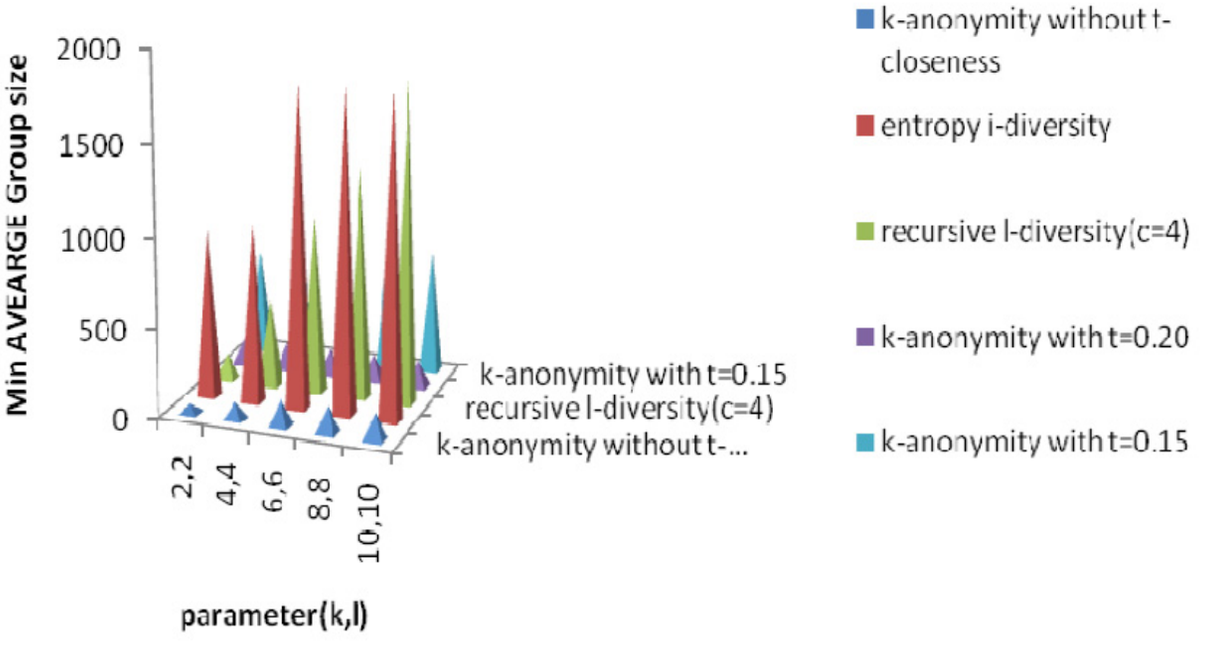

Figure 15. Minimum Average group size Versus Parameter for diverse anonymity techniques

Figure 15. shows Minimum Average group size Versus Parameter for diverse anonymity techniques.

Summarizing the main points we understand that if data utility is of primary concern in a domain, then our rating will be in the order of k-anonymity without t-closeness followed by k-anonymity with t=0.20 followed by k-anonymity with t=0.15 followed by recursive l-diversity(c=4) followed by entropy i-diversity.From the perspective of projection it will be Orthogonal projection followed by sparse projection and at last the Random projection.

Incase of Privacy, is our Spartan we select Random projection as the first choice followed by Orthogonal projection and finally sparse projection. Amongst the different classification algorithms KNN provides better balance between data utility and data privacy than the SVM Classification methods.

The system also caters to the non-functional requirements viz., Reliability, Availability, Maintainability, Portability, Security.

8. Conclusion and future work

In this chapter we propose the new appraised generic scaffold housing the innovative modus operandi for selection of the superlative Anonymity based Privacy Preserving Data Mining (PPDM) Modus Operandi for optimized privacy preserving data mining. We pin our attention on the quandary of rating discretionary concealed data sets on the basis of the values of data privacy and data utility that the Anonymisation produces. Moreover, we have proposed a generic scaffold housing the innovative modus operandi for selection of the superlative anonymisation technique for optimized PPDM; and we have contradistinguished the diverse anonymity techniques and argue solutions to tackle the problems of security threats and attacks in the PPDM in systems.

We (Indumathi, J. et al., (2009)will be able to understand the potential impact of the various Anonymisation based PPDM Techniques only when we model in full-scale applications. We do deem, nonetheless that our argument should make others vigilant to the perils intrinsic to heuristic approaches to Anonymisation based PPDM Limitations. Heuristic methods are based on assumptions which are tacit and not implicit. If for a given data Anonymisation based PPDM limitation problem, the execution of model-based solutions emerges to be too complicated or too costly to carry out, heuristic approaches need to be incorporated with a meticulous analysis aimed at probing the extent to which the approach formalizes rational group inclination structures and/or data user behaviors.

We (Indumathi, J. et al., (2009) have a wealthy plan for future research. We require considering more complicated models of user performance to take into account pragmatic circumstances where numerous users act concurrently and alliances are possible. This will escort logically to such issues as the incorporation of priors and utilities which need special attention.

Author details

J. Indumathi
Department of Information Science and Technology, College of Engineering, Anna University, Chennai, Tamilnadu, India

9. References

A.Machanavajjhala, J.Gehrke, and D.Kifer, et al,(2007) "ℓ-diversity: Privacy beyond k-anonymity", ACM Transactions on Knowledge Discovery from Data (TKDD) Volume 1 Issue 1, March 2007 Article No. 3.

Aggarwal, C. and Yu, P (2004),'A Condensation Approach to Privacy Preserving Data Mining', In Advances in Database Technology - EDBT, pp. 183-199.

Ali Inan, Yucel Saygin, Erkay Savas, Ayca Azgin Hintoglu and Albert Levi (2006), 'Privacy preserving clustering on horizontally partitioned data', Proceedings of the22nd International Conference on Data Engineering Workshops, IEEE.

Atallah, M.J., Bertino, E., Elmagarmid, A.K., Ibrahim, M. And Verykois, V.S. (1999), 'Disclosure limitation of sensitive rules', Proceedings of the IEEE Knowledge and Data Engineering Workshop , IEEE Computer Society Press, Chicago, IL, USA, pp.45–52.

Du, W. and Zhan, Z. (2002), 'Building decision tree classifier on private data', Proceedings of the IEEE ICDM Workshop on Privacy, Security and Data Mining, p.1–8.

Friedman.A, R. Wolff, and A. Schuster, 'Providing k-Anonymity in Data Mining, Int'l J. Very Large Data Bases', vol. 17, no. 4, pp. 789-804, 2008.

Gitanjali, J., Shaik Nusrath Banu, Geetha Mary,A., Indumathi, J. and Uma, G.V. (2007), 'An agent based burgeoning framework for privacy preserving information harvesting systems', Computer Science and Network Security, Vol. 7, No. 11, pp.268–276.

Indumathi J., Uma G.V.(2008), 'A Novel Framework for Optimized Privacy Preserving Data Mining Using the innovative Desultory Technique', International Journal of Computer Applications in Technology ; Special Issue on: "Computer Applications in Knowledge-Based Systems". Vol.35 Nos.2/3/4, pp.194 – 203.

Indumathi, J. and Uma, G.V. (2007 c), 'A new flustering approach for privacy preserving data fishing in tele-health care systems', International Journal on Healthcare Technology and Management (IJHTM) (Special Issue on Tele-Healthcare System Implementation, Challenges and Issues). Vol. 1, Nos. 1/2/3, pp.43–52.

Indumathi, J. and Uma, G.V. (2007a), 'Customized privacy preservation using unknowns to stymie unearthing of association rules', Journal of Computer Science, Vol. 3, No. 12, pp.874–881.

Indumathi, J. and Uma, G.V. (2007b), 'Using privacy preserving techniques to accomplish a secure accord', Computer Scienceand Network Security, Vol. 7, No. 8, pp.258–266.

Indumathi, J. and Uma, G.V. (2008a), 'An aggrandized framework for genetic privacy preserving pattern analysis using cryptography and contravening-conscious knowledge management systems', Molecular Medicine and Advance Sciences, Vol. 4, No. 1, pp.33–40.

Indumathi.J, Dr.G.V.Uma.(2008 c), 'A Panglossian Solitary-Skim Sanitization for Privacy Preserving Data Archaeology',International Journal of Electrical and Power Engineering . Volume 2 Number 3, pp-154 -165, January 2008.

Indumathi.J, K. Murugesan, J.Gitanjali, D. Manjula(2009), "Sprouting Modus Operandi for Selection of the Best PPDM using Agent Based PPDMT in the Health Care Domain," International Journal of Recent Trends in Engineering, Vol. 1, No. 1, pp.627-629May 2009

Jian Yin, Zhi-Fang Tan, Jiang-Tao Ren and Yi-Qun Chen (2005)'An efficient clustering algorithm for mixed type attributes in large dataset', Proceedings of the 4th InternationalConference on Machine Learning and Cybernetics,Guangzhou, Vol. 18–21, pp.1611–1614.

Kantarcioglu, M. and Clifton, C. (2004), 'Privacy-preserving distributed mining of association rules on horizontally partitioned data', IEEE Transactions on Knowledge and Data Engineering, Vol. 16, No. 9, pp.1026–1037.

L. Sweeney, "k-anonymity: a model for protecting privacy", International Journal on Uncertainty, Fuzziness and Knowledge based Systems, 2002, pp. 557-570.

MohammadReza Keyvanpour et al.,(2011),'Classification and Evaluation the Privacy Preserving Data Mining Techniques by using a Data Modification–based Framework,' International Journal on Computer Science and Engineering (IJCSE), Vol. 3, No. 2, pp. 862–870.

N. Li, T. Li, and S. Venkatasubramanian, (2007)"t-Closeness: Privacy Beyond k-anonymity and l-Diversity", In Proc. of ICDE, 2007, pp. 106-115.

Oliveira, S.R.M. and Zaïane, O.R. (2004), 'Privacy preserving clustering by object similarity-based representation and dimensionality reduction transformation', Proceedings of the 2004 ICDM Workshop on Privacy and Security Aspects of Data Mining, pp.40–46.

P. Samarati and L. Sweeney(1998), "Protecting privacy when disclosing information: k-anonymity and its enforcement through generalization and suppression", In Technical Report SRI-CSL-98-04, SRI Computer Science Laboratory, pp.1-191998.

S.V. Iyengar(2002), 'Transforming Data to Satisfy Privacy Constraints', Proc. Eighth ACM SIGKDD, pp. 279-288, 2002.

Vaidya, J. and Clifton, C. (2002), 'Privacy preserving association rule mining in vertically partitioned data', Proceedings of the 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp.639–644.

Vassilios, S.V., Elmagarmid, A., Bertino, E., Saygin, Y. And Dasseni, E. (2004), 'Association rule hiding', IEEE Transactions on Knowledge and Data Engineering, Vol. 16, No. 4, pp. 434–447.

Vasudevan, V., Sivaraman, N., Senthil Kumar, S., Muthuraj, R., Indumathi, J. and Uma, G.V. (2007), 'A comparative study of SPKI/SDSI and K-SPKI/SDSI SYSTEMS', Information Technology Journal, Vol. 6, No. 8, pp.1208–1216.

Verykios V. S., Bertino. E., Fovino I. N., Provenza L. P., Saygin Y.and Theodoridis. Y. (2004 b),'State-of-the-art in privacy preserving data mining',ACM SIGMOD Record, Vol.33, No.1.

Wang, K. and Fung, B. C. M. (2006),". Anonymizing sequential releases,". In Proceedings of the 12th ACM SIGKDD Conference. ACM, New York.

Wang, K., Fung, B. C. M., AND YU, P. S. (2005)," Template-based privacy preservation in classification problems," In Proceedings of the 5th IEEE International Conference on Data Mining (ICDM). 466–473.

Wang, K., Fung, B. C. M., and Yu, P. S.(2007)," Handicapping attacker's confidence: An alternative to k-anonymization,".Knowl. Inform. Syst. 11, 3, 345–368.

Wong, R. C. W., Li., J., Fu, A. W. C., and Wang, K. (2006),". (a,k)-anonymity: An enhanced k-anonymity model for privacy preserving data publishing,". In Proceedings of the 12th ACM SIGKDD. ACM, New York,754–759.

Zhang, Q.,Koudas, N., Srivatsava, D., and Yu, T. (2007),". Aggregate query answering on anonymized tables,". In Proceedings of the 23rd IEEE International Conference on Data Engineering (ICDE).